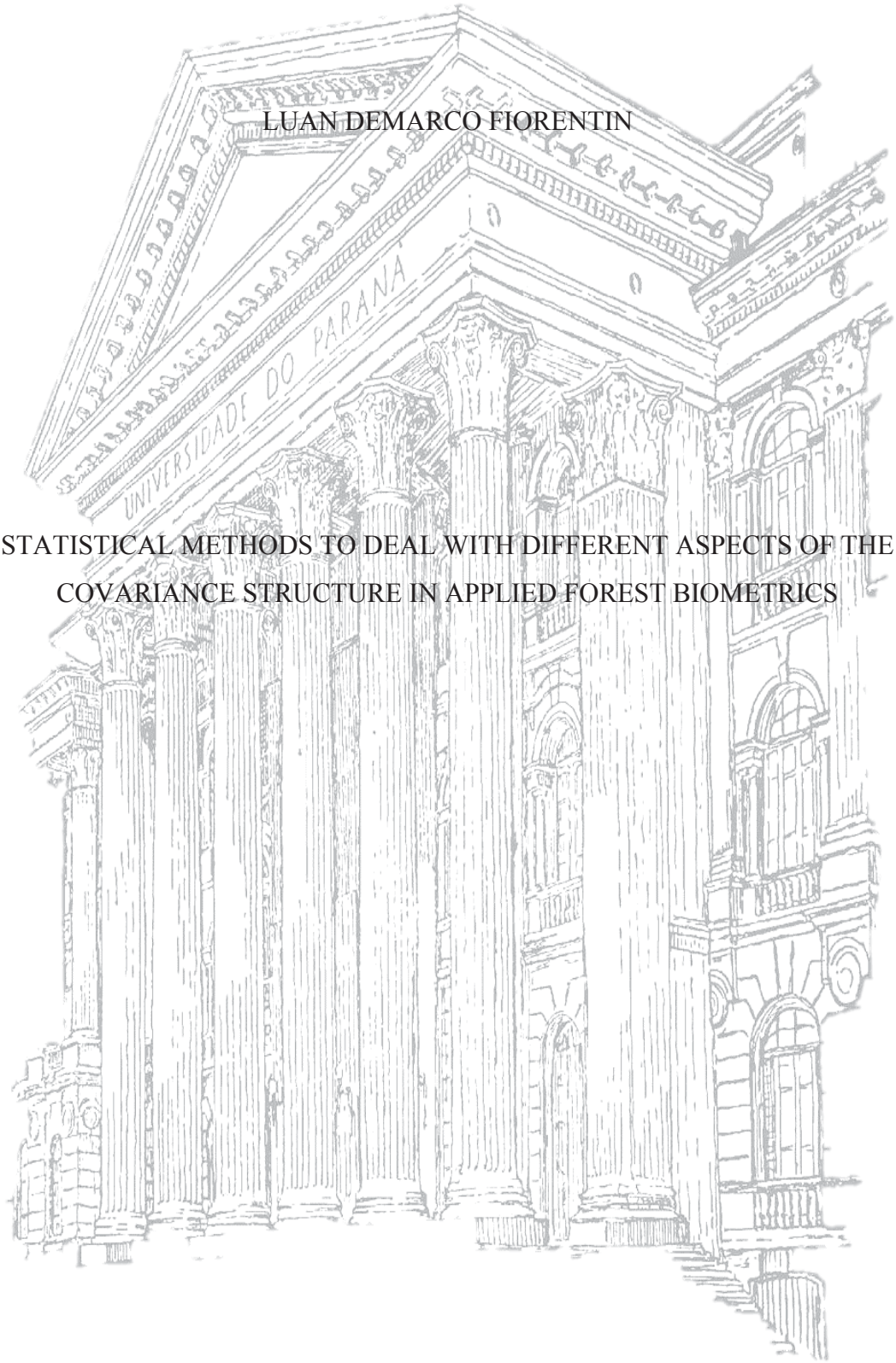
FEDERAL UNIVERSITY OF PARANÁ

LUAN DEMARCO FIORENTIN



STATISTICAL METHODS TO DEAL WITH DIFFERENT ASPECTS OF THE COVARIANCE STRUCTURE IN APPLIED FOREST BIOMETRICS

CURITIBA

2020

LUAN DEMARCO FIORENTIN

STATISTICAL METHODS TO DEAL WITH DIFFERENT ASPECTS OF THE COVARIANCE STRUCTURE IN APPLIED FOREST BIOMETRICS

Tese apresentada como requisito parcial à obtenção do grau de Doutor em Engenharia Florestal, no Curso de Pós-Graduação em Engenharia Florestal, Setor de Ciências Agrárias, Universidade Federal do Paraná.

Orientador:
Professor. Dr. Sebastião do Amaral Machado

Coorientador:
Professor Dr. Wagner Hugo Bonat
Professor Dr. Allan Libanio Pelissari
Professor Dr. Saulo Jorge Téo

CURITIBA

2020

Ficha catalográfica elaborada pela
Biblioteca de Ciências Florestais e da Madeira - UFPR

Fiorentin, Luan Demarco

Statistical methods to deal with different aspects of the covariance structure in applied forest biometrics / Luan Demarco Fiorentin. – Curitiba, 2020.

100 f. : il.

Orientador: Prof. Dr. Sebastião do Amaral Machado
Coorientadores: Prof. Dr. Wagner Hugo Bonat
Prof. Dr. Allan Libanio Pelissari
Prof. Dr. Saulo Jorge Téo

Tese (Doutorado) - Universidade Federal do Paraná, Setor de Ciências Agrárias, Programa de Pós-Graduação em Engenharia Florestal. Defesa: Curitiba, 24/01/2020.

Área de concentração: Manejo Florestal.

1. Florestas - Métodos estatísticos. 2. Biometria. 3. Florestas - Medição - Métodos estatísticos. 4. Teses. I. Machado, Sebastião do Amaral. II. Bonat, Wagner Hugo. III. Pelissari, Allan Libanio. IV. Téo, Saulo Jorge. V. Universidade Federal do Paraná, Setor de Ciências Agrárias. VI. Título.

CDD – 634.9
CDU – 634.0.51

Bibliotecária: Berenice Rodrigues Ferreira – CRB 9/1160

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIENCIAS AGRARIAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ENGENHARIA FLORESTAL - 40001016015P0

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA FLORESTAL da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **LUAN DEMARCO FIORENTIN** intitulada: **STATISTICAL METHODS TO DEAL WITH DIFFERENT ASPECTS OT THE COVARIANCE STRUCTURE IN APPLIED FOREST BIOMETRICS**, sob orientação do Prof. Dr. SEBASTIÃO DO AMARAL MACHADO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.
A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 24 de Janeiro de 2020.

SEBASTIÃO DO AMARAL MACHADO
Presidente da Banca Examinadora

WALMES MARQUES ZEVIANI
Avaliador Externo (DEPARTAMENTO DE ESTATISTICA DA UNIVERSIDADE FEDERAL DO PARANá)

JULIO EDUARDO ARCE
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

JOSE LUIZ PADILHA DA SILVA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

SAULO HENRIQUE WEBER
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

ACKNOWLEDGMENT

To my parents, Amarildo and Angela, and to my brother, Lucas, for supporting me always. You did all this possible with hard work and I thank you a lot for everything. You are the most important people in my life! I love you so much!!!

To my girlfriend, Thaís. You were with me always, even when I was studying in the weekends. You made my days better with your smile!

To my advisor, Dr. Sebastião do Amaral Machado, by the opportunity of studying in a great university and for trusting me over the last years; It is a great honor to known you; you are the biggest name of the Forest Engineering in Brazil. I am very grateful to my first co-advisor, Dr. Allan Libanio Pelissari, for helping me always that I needed; What I can said is that you was not just a co-advisor, you become a great friend, always available for drinking a beer. To my second co-advisor, Dr. Wagner Hugo Bonat, for teaching me about statistics; Your knowledge about statistics inspired me every day, but scared me sometimes too, and I thank you for that; To my third co-advisor, Dr. Saulo Jorge Téo, for teaching me about forest management since my undergraduation.

To my friends Dennis, Gustavo, Igor and Rafael, The Cluster; also known as "canalhas", you did part of my days, sometimes you were helping me and sometimes were stressing me. To my postgraduate's friends; You helped me to know more on forest management; and I will never forget the coffee time talks; I have many friends to thank, but in especial: Ângelo, Antônio, Fran, Gabriel, Hassan, Jonathan, Lucas, Luciano, Rodrigo, Samuel, and Vinicius.

To my LEG's friends Cesar, Elias, Fernando, Guilherme, Henrique, PJ, Vinicius, and Walmes, for making the days more interesting, with a lot of talk about statistics, but always with funny times. I want to thank a lot the Statistical Department of the Federal University of Paraná for give me an opportunity to teach statistics as a teacher assistant.

To Federal University of Paraná and to Postgraduate Program in Forestry Engineering by the opportunity of developing a research in this great university. To National Council of Technological and Scientific Development for supporting me with a PhD research grant.

Luan Demarco Fiorentin.

"Statisticians, like artists, have the bad habit of falling in love with their models"

George Box

RESUMO

A biometria florestal é baseada em pesquisas que visam descrever o comportamento das árvores, com a finalidade de auxiliar o manejo florestal. Nesse contexto, o foco da presente pesquisa foi introduzir modelos estatísticos para melhorar o entendimento quanto ao comportamento das variáveis florestais, combinado com adequadas habilidades de predições. *Capítulo I*: Modelos lineares generalizados de covariância foram introduzidos para modelagem do afilamento do fuste de árvores de *Pinus taeda*. O componente de média foi baseado em um modelo não linear segmentado, enquanto quatro estruturas de covariância foram especificadas para uma explícita modelagem da variância e correlação. Os resultados mostraram que uma matriz de distância Euclidiana e estruturas de médias móveis de ordem 1 a 3 foram adequadas para remover o padrão de variância não constante dos resíduos, bem como a natural autocorrelação entre observações mensuradas ao longo do fuste das árvores. Esses modelos permitiram incluir de forma adequada uma análise de incertezas por meio de intervalos de confiança para a variável resposta. *Capítulo II*: Os modelos lineares generalizados de covariância multivariada foram introduzidos para modelagem conjunta das variáveis respostas altura e volume da *Araucaria angustifolia*, em floresta nativa. As duas variáveis respostas compartilharam informações devido a sua correlação significativa obtida no ajuste multivariado. A função de variância foi um componente importante para a resposta volume e melhorou as estatísticas de ajuste. *Capítulo III*: Métodos de regularização foram introduzidos para selecionar covariáveis correlacionadas em modelos lineares generalizados, para predizer a probabilidade de sobrevivência de árvores de *Pinus taeda*. A função de ligação complemento log-log foi a mais adequada na especificação do modelo Bernoulli. Esse resultado evidenciou que o comportamento da probabilidade de sobrevivência das árvores é assimétrico em relação ao preditor linear. A seleção de covariáveis a partir do procedimento stepwise foi mais parcimoniosa, quando comparada com os métodos de regularização baseados na abordagem elastic net, bem como os casos especiais de penalização lasso e ridge. Os modelos apresentaram ótima habilidade de predição, principalmente para predizer as árvores sobreviventes.

Palavras-chave: Afilamento de fuste. Modelo marginal. Modelo condicional. Sobrevivência. Modelos lineares generalizados.

ABSTRACT

Forest biometrics is based on research that aims to describe the behavior of trees in order to assist the forest management. In this context, our focus was to introduce statistical models able to improve the understanding about the behavior of the forest variables, combined with suitable predictions ability. *Chapter I:* Covariance generalized linear models was introduced for *Pinus taeda* stem taper modeling. We define the mean component based on a non-linear segmented model, while four covariance structures were specified for an explicitly variance and correlation modeling. The results showed that an Euclidean distance matrix and moving average structure of order 1 to 3 were suitable for handling with non-constant variance pattern of the residuals, besides the natural autocorrelation among observations taken over the tree stem. Our models allowed to include a suitable uncertainty analysis based on confidence intervals for the response variable. *Chapter II:* We introduced the multivariate covariance generalized linear models for a jointly fitting of the response variables height and volume of *Araucaria angustifolia*, in native forest. Response variables shared information due to the significant correlation among them on the multivariate fitting. The variance function was an important component for the response volume and have potential to improve the fitting statistics. *Chapter III:* Regularizations methods were introduced for selecting correlated covariates in generalized linear models for predicting the *Pinus taeda* trees survival probability. The complementary log-log link function was the most suitable link function on the specification of the Bernoulli's model. Our result evidenced that the behavior of tree survival probability is asymmetric in relation to the linear predictor. Stepwise procedure was more parsimoniously for selecting covariates, when we compared to the regularization methods based on elastic net approach, as well as the special cases lasso and ridge penalization. Our models presented a great prediction ability, mainly for predicting the survival trees.

Keywords: Stem taper. Marginal model. Conditional model. Survival. Generalized linear models.

SUMMARY

1. GENERAL ASPECTS OF THE THESIS

1.1. GENERAL INTRODUCTION

Forest biometric is a part of forest science focused to obtain current and future yield predictions of forest resources, especially as regards the timber productions for industrial supply. The quantification of individual tree volume and the stem form is a basic procedure for obtaining the financial values of a forest. Thus, some information called of random variables must be collected from the forests, such as the tree diameter, tree height, site occupancy, and many others. These variables are basic components of growth-yield systems, allowing to estimate the changes on a forest and the expected production over the time.

The volume prediction is a fundamental element of the large-scale forest management and planning. The most versatile and accurate approach for estimating tree volume are the modelling methods (DE-MIGUEL et al., 2012), which include a set of statistical approaches with particularly features. The regression methods are largely applied to forest biometric for volume predictions at tree-level. In forestry literature, we can find two main strategies for volume modeling. The first one is based on volume models that describe the total or partial tree stem volume; while the second approach involve the so-called tree stem taper functions. The analytical flexibility afforded by stem taper models has been reported in last decades (WESTFALL & SCOTT, 2010), and can be observed in many research, such as Arias-Rodil et al. (2015), Cao & Wang (2015), Diéguez-Aranda et al. (2006), Fortin et al. (2013), MacFarlane & Weiskittel (2016), De-Miguel et al. (2012), and Sabatia & Burkhart (2015).

Data for the stem taper models usually present interesting features. Due to the natural hierarchical structure of the data, an autocorrelation among observations taken within-tree are expected. This feature is not always well accommodated in taper functions, and the variance of the residuals usually become non-constant over tree stem (LEJEUNE et al., 2009; LI & WEISKITTEL, 2009; MACFARLANE & WEISKITTEL, 2016), which can directly influence the standard errors of the parameter estimates and increase the prediction variance. In this context, we introduce the covariance generalized linear models proposed by Bonat & Jørgensen (2016), which is an alternative approach for stem taper modeling and allows to handle with correlated data in an easy way, besides the non-constant variance. This class of models is based on a marginal model specification and second-moment assumptions, what allows us to obtain a flexible specification of the mean and covariance structures.

Tree height is also an important variable commonly used in growth-yield systems. The diameter at breast height is relatively cheap and can be more accurately measured than the total tree height, usually quantified in a subsample. Height-diameter models are then applied for predicting the height behavior of the trees (CASTEDO DORADO et al., 2006; MEHTÄTALO et al., 2015; TRINCADO et al., 2007). The multivariate case of covariance generalized linear models handle with two or more response variables simultaneously (BONAT, 2017; BONAT et al., 2017). This feature is very interesting, once that we can obtain a jointly model for volume and height predictions and to quantify the correlation between the response variables.

From the height and volume models developed, the total volume of any individual tree is easily obtained in growth-yield systems. However, the natural competition among individuals lead to a fundamental process in forest development related to tree mortality or survival process (SZMYT et al., 2018). The tree survival in forest stands are associated to a set of potential factors (BOSE et al., 2018; ZHANG et al., 2017) that can influence the forest dynamic. The quantification of alive trees indicates the forest structure and how many individuals will be available for industrial supply. In this context, statistical models must be applied for predicting the tree survival probability over the time; and help us to understand how the tree survival manifests in a forest.

Therefore, we present some statistical approaches for handling with correlated data, mainly in the context of tree stem taper modeling; jointly modeling of response variables height and volume; and for tree survival probability in forest stands.

1.2. GENERAL OBJECTIVE

The main objective of this thesis was to present statistical methods with great potential to be applied to forest biometrics. In especial, our focus was to introduce statistical models able to improve the understanding about the behavior of the forest variables at individual tree-level, combined with suitable predictions ability, once that it is a challenge to find models suitable for both description and prediction.

1.3. SPECIFIC OBJECTIVES

We defined the followings specific objectives:

I. To introduce the covariance generalized linear models for *Pinus taeda* L. tree stem taper modeling;
II. To introduce the multivariate covariance generalized linear models for a jointly modeling of height and volume of *Araucaria angustifolia* (Bert.) O. Ktze., in native forest;
III. To present a generalized linear model for *Pinus taeda* L. tree survival modeling.

1.4. THESIS ORGANIZATION

The thesis was structured in three main chapters.

In the first chapter we presented the paper called "*Covariance generalized linear models: an approach for quantifying uncertainty in tree stem taper modeling*". In this research, our focus was to define a suitable covariance matrix based on covariance generalized linear models for tree stem taper modeling.

The second chapter was called "*Joint marginal modeling of height and volume for Araucaria angustifolia*". The idea of this paper was to fit a jointly model for tree height and volume based on multivariate covariance generalized linear model.

The third and last chapter was called "*Generalized linear models for tree survival in loblolly pine plantations*". In this paper, the focus was to select covariates based on stepwise procedure and penalizations methods for fitting a Bernoulli generalized linear model for tree survival.

REFERENCES

ARIAS-RODIL, M.; DIÉGUEZ-ARANDA, U.; PUERTA, F.R.; LÓPEZ-SÁNCHEZ, C.A.; LÍBANO, E.C.; OBREGÓN, A.C.; CASTEDO-DORADO, F. Modeling and localizing a stem taper function for *Pinus radiata* in Spain. **Canadian Journal of Forest Research** , v. 45, p. 647–658, 2015.

BONAT, W.H. Modelling Mixed Types of Outcomes in Additive Genetic Models. **The international journal of biostatistics**, v. 13, n. 2, p. 1–16, 2017.

BONAT, W.H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society. Series C: Applied Statistics**, v. 65, n. 5, p. 649–675, 2016.

BONAT, W.H.; OLIVERO, J.; GRANDE-VEJA, M.; FARFÁN A.; FA, J.E. Modelling the covariance structure in marginal multivariate count models: Hunting in Bioko Island. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 22, n. 4, p. 446–464, 2017.

BOSE, A.K.; WEISKITTEL, A.; KUEHNE, C.; WAGNER, R.G.; TURNBLOM, E.; BURKHART, H.E. Tree-level growth and survival following commercial thinning of four major softwood species in North America. **Forest Ecology and Management**, v. 427, p. 355–364, 2018.

CAO, Q.V.; WANG, J. Evaluation of methods for calibrating a tree taper equation. **Forest Science**, v. 61, n. 2, p. 213–219, 2015.

CASTEDO-DORADO, F.; DIÉGUEZ-ARANDA, U.; ANTA, M.B.; RODRÍGUEZ, M.S.; VON GADOW, K.A generalized height-diameter model including random components for radiata pine plantations in northwestern Spain. **Forest Ecology and Management**, v. 229, n. 3, p. 202–213, 2006.

DE-MIGUEL, S; MEHTÄTALO, L.; SHATER, Z.; KRAID, B.; PUKKALA, T. Evaluating marginal and conditional predictions of taper models in the absence of calibration data. **Canadian Journal of Forest Research**, v. 42, n. 7, p. 1383–1394, 2012.

DIÉGUEZ-ARANDA, U., CASTEDO-DORADO, F.; ÁLVAREZ-GONZÁLEZ, J.G.; ROJO, A. Compatible taper function for Scots pine plantations in northwestern Spain. **Canadian Journal of Forest Research**, v. 36, n. 5, p. 1190–1205, 2006.

FORTIN, M.; SCHNEIDER, R.; SAUCIER, J. Volume and error variance estimation using integrated stem taper models. **Forest Science**, v. 59, n. 3, 2013.

LEJEUNE, G., UNG, C.H.; FORTIN, M.; GUO, X.J.; LAMBERT, M.C.; RUEL, J.C. A simple stem taper model with mixed effects for boreal black spruce. **European Journal of Forest Research**, v. 128, n. 5, p. 505–513, 2009.

LI, R.; WEISKITTEL, A. Development and evaluation of regional taper and volume equations for the primary conifer species in the Acadian Region of North America. **Annals of Forest Science**, v. 67, p. 21–24, 2010.

MACFARLANE, D.W.; WEISKITTEL, A.R. A new method for capturing stem taper variation for trees of diverse morphological types. **Canadian Journal of Forest Research**, v. 46, n. 6, p. 804–815, 2016.

MEHTÄTALO, L.; DE-MIGUEL, S.; GREGOIRE, T.G. Modeling height-diameter curves for prediction. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 826–837, 2015.

SABATIA, C.O.; BURKHART, H.E. On the use of upper stem diameters to localize a segmented taper equation to new trees. **Forest Science**, v. 61, n. 3, p. 411–423, 2015.

SZMYT, J.; TARASIUK, S. Species-specific spatial structure, species coexistence and mortality pattern in natural, uneven-aged Scots pine (*Pinus sylvestris* L.) - dominated forest. **European Journal of Forest Research**, v. 137, n. 1, p. 1–16, 2018.

TRINCADO, G.; VANDERSCHAAF, C.L.; BURKHART, H.E. Regional mixed-effects height-diameter models for loblolly pine (*Pinus taeda* L.) plantations. **European Journal of Forest Research**, v. 126, n. 2, p. 253–262, 2007.

WESTFALL, J.A.; SCOTT, C.T. Taper models for commercial tree species in the northeastern United States. **Forest Science**, v. 56, n. 6, p. 515–528, 2010.

ZHANG, X.; CAO, Q.V.; DUAN, A.; ZHANG J. Modeling tree mortality in relation to climate, initial planting density, and competition in Chinese fir plantations using a Bayesian logistic multilevel method. **Canadian Journal of Forest Research**, v.47, p. 1278–1285, 2017.

2. MODELOS LINEARES GENERALIZADOS DE COVARIÂNCIA: UMA ABORDAGEM PARA QUANTIFICAR INCERTEZAS NA MODELAGEM DO AFILAMENTO DO FUSTE DE ÁRVORES

ABSTRACT

A natural dependence among diameters measured within-tree are expected in taper data, and suitable statistical models are fundamental for analyzing the tree stem form. The main aim of this paper was to introduce the covariance generalized linear models (CGLM) framework in the context of forest biometrics for *Pinus taeda* stem form modeling. The CGLM are based on a marginal specification, which requires a definition of the mean and covariance components. The tree stem mean profiles was modeled by using a non-linear segmented model. The covariance matrix was built considering four strategies of linear combinations of known matrices, which expressed the variance or correlations among observations. Thus, the first strategy (*VarStr*) modeled just the variance of the diameters over the stem as a function of covariates; the correlation among observations was modeled in the second strategy (*CovStr*); the third strategy (*RwStr*) was defined based on a random walk model; the fourth and last strategy (*MmStr*) was based on a structure similar to mixed-effect model, but with a marginal specification. The result showed that the four approaches were quite similar for describing the predicted mean profile. However, differences in the confidence intervals of response relative diameter were quite significant, being directly related to the matrix covariance structures. Marginal and conditional predictions were also performed, and the conditional effects tended to reduce and stabilize the prediction errors over the tree stem. The *CovStr* was the most suitable strategy for modeling the *Pinus taeda* stem taper, due to its robust specification, combined with a suitable prediction ability.

Keywords: Marginal models. Conditional models. Taper functions. Pinus.

COVARIANCE GENERALIZED LINEAR MODELS: AN APPROACH FOR QUANTIFYING UNCERTAINTY IN TREE STEM TAPER MODELING

ABSTRACT

A natural dependence among diameters measured within-tree are expected in taper data, and suitable statistical models are fundamental for analyzing the tree stem form. The main aim of this paper was to introduce the covariance generalized linear models (CGLM) framework in the context of forest biometrics for *Pinus taeda* stem form modeling. The CGLM are based on a marginal specification, which requires a definition of the mean and covariance components. The tree stem mean profiles was modeled by using a non-linear segmented model. The covariance matrix was built considering four strategies of linear combinations of known matrices, which expressed the variance or correlations among observations. Thus, the first strategy (*VarStr*) modeled just the variance of the diameters over the stem as a function of covariates; the correlation among observations was modeled in the second strategy (*CovStr*); the third strategy (*RwStr*) was defined based on a random walk model; the fourth and last strategy (*MmStr*) was based on a structure similar to mixed-effect model, but with a marginal specification. The result showed that the four approaches were quite similar for describing the predicted mean profile. However, differences in the confidence intervals of response relative diameter were quite significant, being directly related to the matrix covariance structures. Marginal and conditional predictions were also performed, and the conditional effects tended to reduce and stabilize the prediction errors over the tree stem. The *CovStr* was the most suitable strategy for modeling the *Pinus taeda* stem taper, due to its robust specification, combined with a suitable prediction ability.

Keywords: Marginal models. Conditional models. Taper functions. Pinus.

2.1. INTRODUCTION

Modeling forest variables is one of the most important goals of forest research. Variables which are frequently modeled include tree volume (BOSE et al., 2018), tree height-diameter relationship (MACPHEE et al., 2018), tree diameter growth (SHARMA et al., 2017), individual tree basal area increment (TENZIN et al., 2017), tree dominant height growth (SEKI & SAKICI, 2017) and tree stem form (ARIAS-RODIL et al., 2015a; GÓMEZ-GARCÍA et al., 2013; WESTFALL et al., 2016; WESTFALL & SCOTT, 2010). Other interesting research involving forest models can be found in Gomat et al. (2011), Mäkinen et al. (2018), Nascimento et al. (2014), Riofrío et al. (2017) and Sharma & Reid (2017).

The tree stem form modeling has special importance in the context of forest management for estimating the individual total volume and multiple timber products. The changes of the diameter along the bole is a function of the tree diameter and height, being generally designated as taper function (BURKHART & TOMÉ, 2012). The flexibility of taper functions provides additional indirect estimates such as: I) total stem volume; II) diameter at any point along the stem; III) merchantable volume and merchantable height to any top diameter and from any stump height; and IV) individual log volumes of any length at any height (KOZAK, 2004).

The stem form modeling requires multiple diameter and height measures within an individual tree. As consequence, autocorrelation among observations taken within-tree are expected due to the natural hierarchical structure of the data. Lejeune et al. (2009) mentioned that over the last decades, the mixed-effects model has become popular in forestry literature for analyzing stem taper data. The mixed-effects models have been a statistical approach widely used because enables to estimate the between-tree and within-tree variability using fixed-effects and random-effects parameters (ARIAS-RODIL et al., 2015b). Besides, the random-effects parameters of mixed models are also frequently used for quantifying at least partly of the autocorrelation between observations. However, Li & Weiskittel (2009) highlighted that when the correlations are not fully eliminated, and heterogeneous variance are observed, covariance structures can be included, jointly with variance functions.

Even the variance and covariance structures play a key role in the models for correlated data, these structures have not been widely explored in taper functions. Research in tree stem taper indicated that when the random-effects do not fully eliminated the autocorrelation within-tree, a covariance structure must be included in the model, being frequently restricted to the first-order autoregressive or first and second-order continuous autoregressive covariance

structure and four-banded toeplitz (ARIAS-RODIL et al., 2015a; DIÉGUEZ-ARANDA et al., 2006; FORTIN et al., 2013; GÓMEZ-GARCÍA et al., 2013; LEJEUNE et al., 2009; LI & WEISKITTEL, 2009; MACFARLANE & WEISKITTEL, 2016; SABATIA, 2016; YANG et al., 2009). On the other hand, the residual heteroscedasticity within-tree is commonly modeled by the power-of-the-mean variance function, variance power weighting structure, or an exponential variance function (LEJEUNE et al., 2009; LI & WEISKITTEL, 2009; MACFARLANE & WEISKITTEL, 2016). These modeling approaches also require selecting the best combination of random-effects parameters that contributed most to unexplained variation of the response variable (MACFARLANE & WEISKITTEL, 2016), what can be a laborious and time demanding process. Besides, the short list of prespecified covariance structures to handle with taper models in specialized statistical software is a limiting factor.

The mixed-effect models are based on a conditional specification, which implies that the distribution of response variable is conditioned to the random-effects. In some context, as the quantitative genetic analysis, Bonat (2017) mentioned that inconvenient features are obtained from conditional specification when genetic additive effects are being evaluated in hypothesis testing. The marginal distribution of the response variable usually cannot be obtained by closed-form from the conditional specification. The author still highlighted that the model parameters must be carefully interpreted, once that the covariate effects are conditional on the random-effects, while the covariance structure is marginal for the random-effects rather than for the outcome. In contrast to the conditional models there are the marginal models, which are obtained from a marginal specification of the expected value of response variable. Lee & Nelder (2004) explained the main distinction between marginal and conditional models have often been related whether the parameters are to describe an individual´s response or the marginal mean response to changing covariates. Therefore, the main advantage of marginal models is allowing a direct inference for the response variable, such as the evolution of population average response and associations (VERBEKE et al., 2014).

Recently, Bonat & Jørgensen (2016) developed a general modeling framework called covariance generalized linear models (CGLM). The CGLM is quite flexible for modeling univariate (UCGLM) and multivariate (MCGLM) correlated data, considering response variable of mixed types, and allows to define many covariance structures for repeated measures, longitudinal, spatial and spatio-temporal data. This modeling approach is based on a marginal model specification and second-moment assumptions, what allows to introduce a flexible specification of the mean and covariance structures. Besides easily handle with univariate and multivariate response variable, the CGLM framework can introduce explicitly a covariance

structure using a linear combination of known matrices, such as the moving average model, unstructured matrix, inverse of Euclidean distance, and compound symmetry structures (BONAT et al., 2017; BONAT & JØRGENSEN, 2016).

The uncertainty in model predictions has been studied in the last years in a large variety of contexts (BERGER et al., 2014; FORTIN et al., 2016; MANSO et al., 2018; OIJEN, 2017). McRoberts & Westfall (2014) divided the sources of uncertainty due to the model misspecification; uncertainty in observed values of the covariates; residual variability from correctly specified models; and the uncertainty in the model parameter estimates. The structure of the CGLM allows to quantify the uncertainty related to the predictions of response variables considering a flexible covariance matrix. In this research, the focus was to quantify the uncertainty related to the natural variability of the response variables by specifying a suitable covariance matrix.

The CGLM approach is quite promising and presents a great potential in forest modeling, once that quantitative forest variables usually present correlated data, high and non-constant variance for the response variable, as can be usually observed in tree stem taper data (ARIAS-RODIL et al., 2015a; DIÉGUEZ-ARANDA et al., 2006; YANG et al., 2009). Therefore, our research hypothesis is that the covariance generalized linear models will be suitable for modeling the behavior of tree stem taper.

Due to the importance of tree stem form modeling in the forest management, the aim of this paper was to introduce the covariance generalized linear model framework in the context of forest biometrics. Specific objectives were I) to analyze the fit of a segmented nonlinear model for response variable relative diameter using covariance generalized linear models; II) to propose a suitable marginal covariance matrix as a linear combination of known matrices to take into account the non-constant pattern of variance and correlation among observations of the response variable; III) to generate conditional predictions for increasing the prediction ability of the marginal models.

2.2. MATERIAL AND METHODS

2.2.1. DATA SET

We obtained a cross-sectional data from loblolly pine (*Pinus taeda* L.) forest stands established in Midwest region of the Santa Catarina State, Brazil. Forest plantation area covers about 5,786.75 hectares distributed in 164 stands. The taper data set was obtained by measuring

427 samples trees, which were randomly selected in unthinned and thinned stands covering the range of ages.

The random variable diameter at breast height (DBH, in cm) outside bark were directly measured on each tree at 1.3 m of height. The trees were felled and the random variable total height (H, in m) was also directly measured. We took 16 repeated measures of random variable diameter (d, in cm) outside bark at 0%, 0.5%, 1%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of total height. Smalian's method was used to calculate the log volume for each section. The tree-top volume was calculated based on cone's formula. For each sample tree, the random variable individual total volume (V, in m^3) outside bark was obtained by summing the partial volume of each section.

For a detailed exploratory data analysis, the dataset was split in four age classes according to thinning level: C1 was the first age class and the individuals are 4 to 7 years old, none thinning was applied; individuals in the C2 age class are 8 to 11 years old, but one thinning from below plus systematic (in the seventh line) were apply, by removing 50% of the trees per hectare; individuals in the C3 age class are 12 to 19 years old, and two thinning from below were applied, by removing 40% of the remaining trees; C4 was the last age class and the individuals are 20 to 30 years old, three thinning from below were applied, by removing 30% of the remaining trees.

2.2.2. MARGINAL SPECIFICATION OF THE COVARIANCE GENERALIZED LINEAR MODEL

In the context of covariance generalized linear model (CGLM), a general formulation for expected value and variance for the response variable is given as

$$\mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu} = \mathrm{g}^{-1}(\mathbf{X}\boldsymbol{\beta}) \tag{1}$$

$$\mathrm{Var}[\mathbf{Y}] = \boldsymbol{\Sigma} = \mathrm{V}(\boldsymbol{\mu}; \mathbf{p})^{1/2}\Omega(\boldsymbol{\tau})\mathrm{V}(\boldsymbol{\mu}; \mathbf{p})^{1/2}, \tag{2}$$

where $\mathbf{Y}$ is an $\mathrm{N} \times 1$ response vector, being N the number of total observation; $\mathbf{X}$ is an $\mathrm{N} \times \mathrm{K}$ design matrix, being K the number of covariates; $\boldsymbol{\beta}$ is an $\mathrm{K} \times 1$ regression parameters vector; g is the differentiable and monotonous link function; $\mathrm{V}(\boldsymbol{\mu}; \mathbf{p}) = \mathrm{diag}\{\vartheta(\boldsymbol{\mu}; \mathbf{p})\}$ is a diagonal matrix whose main entries are given by the variance function $\vartheta(\boldsymbol{\mu}; \mathbf{p})$ applied elementwise to

the vector $\boldsymbol{\mu}$; $\mathbf{p}$ is a vector of power parameter; $\Omega(\boldsymbol{\tau}) = \tau_0\mathbf{Z_0} + \cdots + \tau_T\mathbf{Z_T}$, Z_t with $t = 0, \ldots, K$ are known matrices reflecting the covariance structure; and $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_T)$ is a $(T + 1) \times 1$ dispersion parameter vector.

In our approach, different assumptions about the response variable distribution can be performed for different choices of the variance function, similar that occurs with the generalized linear models. The power variance functions $\vartheta(\boldsymbol{\mu}; p) = \boldsymbol{\mu}^p$ characterizes the Tweedie family of distributions and the most important special cases are covered Normal ($p = 0$), Poisson ($p = 1$), Gamma ($p = 2$) and Inverse Gaussian ($p = 3$) distributions.

The mean structure is called linear or non-linear predictor, and we considered just an identity link function. The assumption of independent observations appears in the covariance matrix of the equation (2). For introducing some dependence structure between observations, we specify the $\Omega(\boldsymbol{\tau})$ as a non-diagonal matrix. The dispersion matrix $\Omega(\boldsymbol{\tau})$ describes the dependence within response variable and does not depend on the mean structure. This approach is similar to the idea of a working correlation matrix in the generalized estimation equation. The CGLM approach is different because it is proper to model $\Omega(\boldsymbol{\tau})$ in terms of a linear combination of known matrix (BONAT & JØRGENSEN, 2016). This structure is called as matrix linear predictor and it is interpreted analogue of the linear predictor of the mean structure.

The second-order specification requires a non-linear predictor and a linear covariance matrix. The CGLM allows us to divide the set of parameters into two subsets $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\lambda})^T$. The $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_R)^T$ denote an $R \times 1$ vector containing the regression parameters and $\boldsymbol{\lambda} = (\tau_0, \ldots, \tau_T)^T$ denote a $(T + 1) \times 1$ vector of dispersion parameters. The quasi-score function perform estimates of the regression parameters, while the dispersion parameters are estimated by the Pearson estimating function. Inferences about the parameter estimates are based on asymptotic distribution of the parameters vector. For more details on the methods see Bonat (2018) and Bonat & Jørgensen (2016).

The covariance generalized linear models were fitted in the mcglm package (BONAT, 2018), available on the R software (R CORE TEAM, 2019). Besides fitting the models, many auxiliary functions are implemented for building the components of the matrix linear predictor, which are detailed in the covariance structure subsection. The package uses the modified chaser algorithm for obtaining the estimates of the model parameters. Furthermore, the reciprocal likelihood algorithm was implemented with an additional tuning for controlling the step length.

2.2.3. MEAN STRUCTURE

Taper models must have the important feature of high flexibility for describing the non-linear nature of tree tapering with different degrees of curvature over stem (KUBLIN et al., 2013). In this sense, we selected the non-linear segmented model introduced by Max & Burkhart (1976) for describing the tree stem taper. The model is composed by three polynomial submodels linked by two inflexion points. The flexibility of this segmented model in describing complex tree stem form has been previously explored by Arias-Rodil et al. (2017), Cao & Wang (2015), Diéguez-Aranda et al., (2006), MacFarlene & Weiskittel (2016) and Sabatia & Burkhart (2015). The model is given as

$$\mathrm{E}[\mathbf{Y}] = \beta_1(\mathbf{X} - 1) + \beta_2(\mathbf{X^2} - 1) + \beta_3(\alpha_1 - \mathbf{X})^2\dot{\mathrm{I}}_1 + \beta_4(\alpha_2 - \mathbf{X})^2\dot{\mathrm{I}}_2, \quad (4)$$

where $\mathbf{Y} = \mathbf{dD^{-1}}$ is a vector of the response variable relative diameter; $\mathbf{d}$ is a vector of diameters measured over the tree stem; $\mathbf{D}$ is a vector of diameters at breast height measured in each tree; $\mathbf{X} = \mathbf{hH^{-1}}$ is a vector of predictor variable relative height; $\mathbf{h}$ is a vector of partial heights measured over the tree stem; $\mathbf{H}$ is a vector of total heights; $\boldsymbol{\alpha}_s$ are the inflexion points to be estimated ($s = 1, 2$); $\boldsymbol{\beta}_f$ are the parameters to be estimated ($f = 1, 2, 3, 4$); $\dot{\mathrm{I}}_q = 1$ if $\mathbf{X} \leq \alpha_s$ and 0 otherwise, which are a dummy indicator variable vector; $g(\cdot)$ is an identity link function.

The link functions are a fundamental component of the CGLM, once they link the expectation of the response variable with the covariates. The mcglm package has a set of default link functions which are suitable for many types of covariates. However, our model is a non-linear function $f(\cdot)$ of the parameter vector, $\boldsymbol{\theta} = [\beta_1\ \beta_2\ \beta_3\ \alpha_1\ \beta_4\ \alpha_2]^{\mathrm{T}}$, and a proper connection between the non-linear predictor and the response is required. Thus, we specified a component to connect the mean model and the response based on partial derivatives of the parameter vector. These functions were implemented in R language and can be easily used on the package. The expressions of the partial derivatives are given as

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X})}{\partial \beta_1} = (\mathbf{X} - 1);$$

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X}_{ij})}{\partial \beta_2} = (\mathbf{X^2} - 1);$$

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X})}{\partial \beta_3} = (\alpha_1 - \mathbf{X})^2 \dot{\mathrm{I}}_1;$$

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X})}{\partial \alpha_1} = 2\beta_3(\alpha_1 - \mathbf{X})\dot{\mathrm{I}}_1;$$

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X})}{\partial \beta_4} = (\alpha_2 - \mathbf{X})^2 \dot{\mathrm{I}}_2;$$

$$\frac{\partial f(\boldsymbol{\theta}, \mathbf{X})}{\partial \alpha_2} = 2\beta_4(\alpha_2 - \mathbf{X})\dot{\mathrm{I}}_2.$$

2.2.4. COVARIANCE STRUCTURE

The trees in forest stands are usually correlated in some way. Thus, it is expected that close individuals in space present higher association degree. This natural feature has potential to influence the tree stem form due to the intra-specific competition. In forest literature, many research aims to estimate the mean behavior of tree stem profile, which is a stochastic process where the observations taken within-tree are not independent. However, the focus of this research is understanding how the correlation pattern between relative diameters within-individuals occurs in *Pinus taeda* trees.

The second-moment assumption of the CGLM requires the specification of the expectation and a matrix linear predictor. The expectation model was already defined by a non-linear segmented model. However, the non-constant variance and the higher correlation values among diameters taken within-tree motivated us to develop strategies for building the matrix linear predictor. In this sense, our main interest was to present four new approaches for modeling the covariance matrix in the context of tree stem taper analysis.

Strategy 1 – VarStr: we modeled the variance of response variable only based on covariates of easy access. Then, components of the matrix linear predictor were specified without incorporating the repeated measures structure. The variance structure was directly modeled based on the main effects of the covariates relative height ($\mathbf{H_r}$) and age ($\mathbf{A}$), besides their second order terms $\mathbf{H_r^2}$ and $\mathbf{A}^2$, and interaction effects between $\mathbf{H_r}$: $\mathbf{A}$ and $\mathbf{H_r^2}$: $\mathbf{A}^2$. Identity matrix ($\mathbf{I}$) was a pre-specified component independent of the others two covariates an its

transformations. For clarity, consider a particular group with the first three observations, the components of the matrix linear predictor based on the covariates were given as

$$\mathbf{H_r} = \begin{bmatrix} H_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{ij} \end{bmatrix};$$

$$\mathbf{A} = \begin{bmatrix} A_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_{ij} \end{bmatrix};$$

$$\mathbf{H_r^2} = \begin{bmatrix} H_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{ij}^2 \end{bmatrix};$$

$$\mathbf{A^2} = \begin{bmatrix} A_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_{ij}^2 \end{bmatrix};$$

$$\mathbf{H_r}\mathbf{:}\,\mathbf{A} = \begin{bmatrix} H_{11}A_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{ij}A_{ij} \end{bmatrix};$$

$$\mathbf{H_r^2}\mathbf{:}\,\mathbf{A^2} = \begin{bmatrix} H_{11}^2A_{11}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H_{ij}^2A_{ij}^2 \end{bmatrix};$$

where H_{ij} is the j-th relative height of the i-th individual; and A_{ij} is the j-th age of the i-th individual, being $A_{ij} = A_i, \forall_j$.

Strategy 2 – CovStr: we defined the components of matrix linear predictor considering just the correlation structure among observations of response variable. The moving average model of order p ($\mathbf{MA}(p)$) was specified, and we tested the order terms ranging from 1 to 10. We also build components based on the inverse of Euclidean distance between pairs of relative heights ($\mathbf{ED_h}$) and between pairs of observations ($\mathbf{ED_o}$). Identity matrix ($\mathbf{I}$) was pre-specified independent of the twelve tested structures. The components of the matrix linear predictor associated with $\mathbf{MA}(1)$, $\mathbf{MA}(2)$, $\mathbf{ED_h}$ and $\mathbf{ED_o}$ structures were given as

$$\mathbf{MA}(1) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix};$$

$$\mathbf{MA}(2) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix};$$

$$\mathbf{ED_{hi}} = \begin{bmatrix} 0 & \cdots & \mathrm{dh}_{1j'}{}^{-1} \\ \vdots & \ddots & \vdots \\ \mathrm{dh}_{j1}{}^{-1} & \cdots & 0 \end{bmatrix};$$

$$\mathbf{ED_{oi}} = \begin{bmatrix} 0 & \cdots & \mathrm{do}_{1j'} \\ \vdots & \ddots & \vdots \\ \mathrm{do}_{j1} & \cdots & 0 \end{bmatrix}^{-1};$$

where dh_{ij} is the Euclidean distance between the j-th and i-th relative height of the same tree; do_{ij} is the Euclidean distance between the j-th and i-th position of relative height of the same tree.

Strategy 3 – RwStr: we proposed to model the matrix linear predictor as a random walk model. This structure is frequently used for analyzing time series and spatial data; however, it was not explored for modeling tree stem data yet. Also called as precision matrix, the model is specified by the inverse of the dispersion matrix in the following way

$$\boldsymbol{\Omega}(\delta, \rho)^{-1} = \delta(\mathbf{B} - \rho\mathbf{W}),$$

where **W** is a neighborhood matrix, and in the context of stem taper the neighborhoods means the pairs of observations that were taken in a sequence; **B** is a diagonal matrix with the number of neighborhoods in the main diagonal; δ is a precision parameter; and ρ is a spatial autocorrelation parameter. Different from the other strategies where it were applied just an identity covariance link function, we proposed this equation as a linear covariance model using the inverse covariance link function (BONAT & JØRGENSEN, 2016). The formulation was given as

$$\boldsymbol{\Omega(\tau)^{-1}} = \tau_0\mathbf{B} + \tau_1\mathbf{W},$$

where $\tau_0 = \delta$; and $\tau_1 = -\delta\rho$. The components of the matrix linear predictor for the first three observations taken within-tree were given as

$$\mathbf{W} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix};$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

The spatial autocorrelation parameter was still used as an estimate of correlation between neighborhoods (relative diameters) and was given as

$$\hat{\rho} = \hat{\tau}_1\hat{\tau}_0^{-1}.$$

The estimates of variance of the spatial autocorrelation parameter was computed by delta method, which is a general method for approximating the variance of a function of random variables with normal distribution and known covariance matrix. Confidence intervals for spatial autocorrelation parameter was obtained as

$$\mathrm{CI}(\rho, \gamma) = \hat{\rho} \pm \mathrm{Z}_{\gamma/2}\sqrt{\widehat{\mathbf{V}}(\hat{\rho})},$$

where $\hat{\rho}$ is an estimated parameter of spatial autocorrelation; $\mathrm{Z}_{\gamma/2}$ is a quantile of normal distribution for γ confidence level; $\widehat{\mathbf{V}}$ is an estimated variance matrix, computed as $\widehat{\mathbf{V}} = \mathbf{J}\hat{\mathbf{C}}\mathbf{J}^{\mathrm{T}}$; $\mathbf{J}$ is a derivative matrix of dispersion parameters $\hat{\tau}_0$ and $\hat{\tau}_1$; and $\hat{\mathbf{C}}$ is a covariance matrix of the fitted model.

Strategy 4 – MmStr: a common way to fit taper function is based on conditional specification of a non-linear mixed-effect model, considering a normal distribution of response variable. However, in a similar design of mixed model, we presented the marginal specification for taking into account the repeated measures effects within-tree for the covariates

$$Z_1 = (\mathbf{X} - 1)$$

$$Z_2 = (\mathbf{X^2} - 1).$$

Thus, the covariance structure of the mixed-effect model is a special case of our approach.

Components of the matrix linear predictor for variance structures based on covariates, distance structures, moving average model, random walk structures, marginal model and identity matrix were created using the auxiliary functions *mc_dglm*, *mc_dist*, *mc_ma*, *mc_rw*, *mc_mixed* and *mc_id*, respectively, obtained from mcglm package (BONAT, 2018) on the R statistical software (R CORE TEAM, 2019).

2.2.5. UNCERTAINTY IN THE PREDICTIONS

The uncertainty analysis in tree stem taper models is a fundamental topic for evaluating the errors in the statistical models. In this research, our focus was to quantify the uncertainty associated to the response variable relative diameter. Thus, confidence intervals for the response were computed for each modeling strategy by the expression given as

$$\mathrm{CI}(\mathbf{y}, \gamma) = \hat{\boldsymbol{\mu}} \pm Z_{\gamma/2}\sqrt{\hat{\mathbf{C}}},$$

where $\hat{\boldsymbol{\mu}}$ is an $N \times 1$ vector of estimated value; $Z_{\gamma/2}$ is a quantile of normal distribution for γ confidence level; and $\hat{\mathbf{C}}$ is an estimated main diagonal of covariance matrix $\boldsymbol{\Sigma}$ of the fitted values.

2.2.6. MODEL SELECTION

Due to the large number of potential components for the matrix linear predictor in *VarStr* and *CovStr* strategies, in this section we introduce a variable selection criterion for selecting a suitable matrix linear predictor for taper functions. The score information criterion (SIC) was initially proposed by Stoklosa et al. (2014) for selecting covariates in linear predictor for generalized estimating equations. However, Bonat et al. (2017) extended the SIC for selecting components of the matrix linear predictor in the context of MCGLM. These authors

highlighted that the main idea behind SIC is to use the generalized score statistics as a quadratic approximation to the log-likelihood ratio statistics like an information criterion. The main advantage of this approach is once the SIC is a function of the parameters vector $\boldsymbol{\tau}$, only the null model needs to be fitted and the SIC can be computed for all candidate models without actually fitting them.

For selecting a suitable matrix linear predictor, we defined the following steps: 1) to define all the candidates components for the matrix linear predictor; 2) to fit the model considering just the identity covariance matrix, i.e., to fit a simple intercept model, also called as null model; 3) to compute the SIC for all candidates, and select that one with the lowest value for SIC; 4) to update the candidate components, and to fit the model considering the identity covariance matrix plus the component selected in previous step; 5) repeat step three and four until the chi-square test indicates a non-significance for the components. This procedure was performed using the *mc_sic_covariance* function from the mcglm package (BONAT, 2018) on the R statistical software (R CORE TEAM, 2019).

Due to the different number of estimated parameters in each strategy we developed, two information criterions were used as goodness-of-fit statistics to compare the models. Initially, we calculated the Gaussian log-likelihood (logLik) measure given by

$$\text{logLik}(\boldsymbol{\theta}) = -\frac{\text{N}}{2}\log(2\pi) - \frac{1}{2}\log\left|\widehat{\mathbf{C}}\right| - (\mathbf{Y} - \widehat{\boldsymbol{\mu}})^{\text{T}}\widehat{\mathbf{C}}^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}),$$

where $\boldsymbol{\theta}$ is a vector of estimated parameters; N is the total number of observations; $\mathbf{Y}$ is an $\text{N} \times 1$ vector of observed value; $\widehat{\boldsymbol{\mu}}$ is an $\text{N} \times 1$ vector of expected value; and $\widehat{\mathbf{C}}$ is a covariance matrix of the fitted model. Bonat (2018) combined penalty terms with Gaussian log-likelihood to obtain the Akaike (AIC) and Bayesian (BIC) information criterion. Thus, the Akaike and Bayesian information criterion were respectively given by

$$\text{AIC}(\boldsymbol{\theta}) = 2(\text{R} + \text{S}) - 2\text{loglik}(\boldsymbol{\theta}),$$

$$\text{BIC}(\boldsymbol{\theta}) = \log\text{N}(\text{R} + \text{S}) - 2\text{loglik}(\boldsymbol{\theta}),$$

where R is the number of parameters; and S is the number of dispersion parameters. Still, we calculated the mean squared error of the predictions (MSE) as precision measure given by

$$\text{MSE} = \sum_{\text{i=1}}^{\text{N}} N^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}})^2.$$

We also performed a graph analysis about the Pearson's residual (PR), given by

$$\text{PR} = \widehat{\mathbf{C}}^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}).$$

2.2.7. CONDITIONAL PREDICTION OF RESPONSE VARIABLE

As we previously mentioned, the covariance generalized linear models are based on marginal specification of response variable. Thus, the parameter interpretations are performed just over the population mean. This means that the parameters describe the behavior of the mean population to a change in a covariate without considering the heterogeneity among subjects. However, conditional models account part of total unknown variance among subjects by including terms with random-effects, as the mixed-effects models, and usually performed better for predicting the response variable.

In this sub-section, our main objective was increasing the prediction ability to predict individual response of our developed models. Thus, conditional predictions of response variable relative diameter were obtained from equation (5) when we conditionate the fitted value on the $(\text{j}+1)$th relative height given the fitted value on the (j)th position of the (i)th tree. Due to the reason that the observations are not independent within-tree, we considered the covariance matrix in the predictions as fallow

$$\widetilde{\boldsymbol{\mu}}_{\mathbf{i(j+1)|j}} = \widehat{\boldsymbol{\mu}}_{\mathbf{i(j+1)|j}} + \widehat{\mathbf{C}}_{\mathbf{(j+1),j}}\,\widehat{\mathbf{C}}_{\mathbf{j,j}}^{-\mathbf{1}}(\mathbf{y}_{\mathbf{ij}} - \widehat{\boldsymbol{\mu}}_{\mathbf{ij}}), \qquad (5)$$

where $\widetilde{\boldsymbol{\mu}}$ is a vector of conditional predictions of response variable; $\widehat{\boldsymbol{\mu}}$ is a vector of marginal prediction of response variable; $\mathbf{y}$ is a vector of observed response variable; and $\widehat{\mathbf{C}}$ is a covariance matrix of the fitted model. For clarity, suppose that we fitted a value for 50% relative height, then we conditionate the prediction to the fitted value at previously measure on the 40% relative height. However, we did not perform conditional predictions for fitted values on the first (0%) and last (100%) relative height measured on the tree stem.

Conditional predictions performed for individuals used for fitting data was compared with marginal predictions. The analysis was based on mean squared error and the bias over the stem. The *VarStr* strategy was not considered in this sub-section, once that the model did not present any covariance parameter in the specification of the matrix linear predictor.

$$\text{Bias} = \sum_{i=1}^{N} N^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}).$$

The last expression, equation (5), directly depends on observations of diameters taken over the tree. However, diameter at breast height usually is the diameter measured in the forest inventories. Thus, for practical purpose, conditional predictions were also performed for new individuals, where the model was only conditioned to the measures taken at the diameter at breast height (**d**), independent of which relative height we used for predicting the diameter.

2.3. RESULTS

In this section, we present a brief description about the covariates of the data set and the behavior of response variable relative diameter over time. We also specified the non-linear predictor and the matrix linear predictor for each modeling approach. Additionally, we applied the covariance generalized linear model for a marginal modeling of the stem taper over time. Lastly, we performed predictions of relative diameters considering a conditional specification of our marginal models.

2.3.1. EXPLORATORY DATA ANALYSIS

For a detailed analysis of *Pinus taeda* sample trees, we performed the scatterplots represented in FIGURE 2.1 for the covariates diameter, height and individual volume, besides the boxplots of this variables by age class where we can notice their dynamic over time. A right asymmetric distribution with non-constant variance pattern for the response variable was empirically observed in FIGURE 2.2. The behavior of relative diameter changed according to the variables relative height and age class, indicating a possible interaction between them. Furthermore, higher asymmetry and variance were observed for lower relative heights, which were more expressive for younger trees (C1 and C2) than older trees (C3 and C4). Still, it was clear that trees belonging to C1 and C2 classes had their stem form more conical than trees from

C3 and C4 classes. This brief exploratory analysis also suggested that the variable relative height and age should be included in taper models as covariates for variance modeling.

FIGURE 2.1 - SCATTERPLOT OF THE DIAMETER, HEIGHT AND VOLUME, AND BOXPLOT BY AGE CLASS

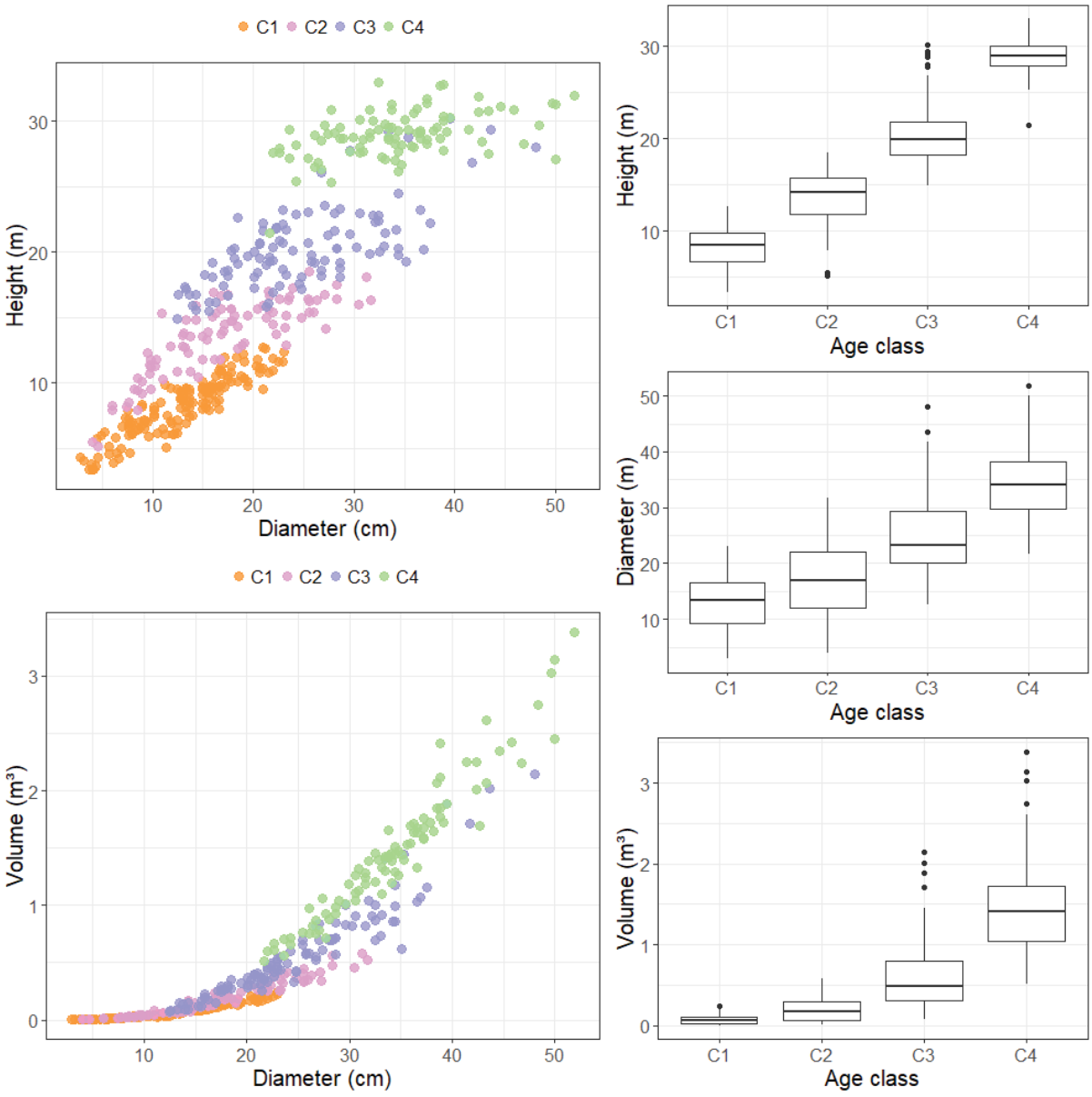


FIGURE 2.2 - BOXPLOT OF RESPONSE VARIABLE RELATIVE DIAMETER BY RELATIVE HEIGHT AND AGE CLASS

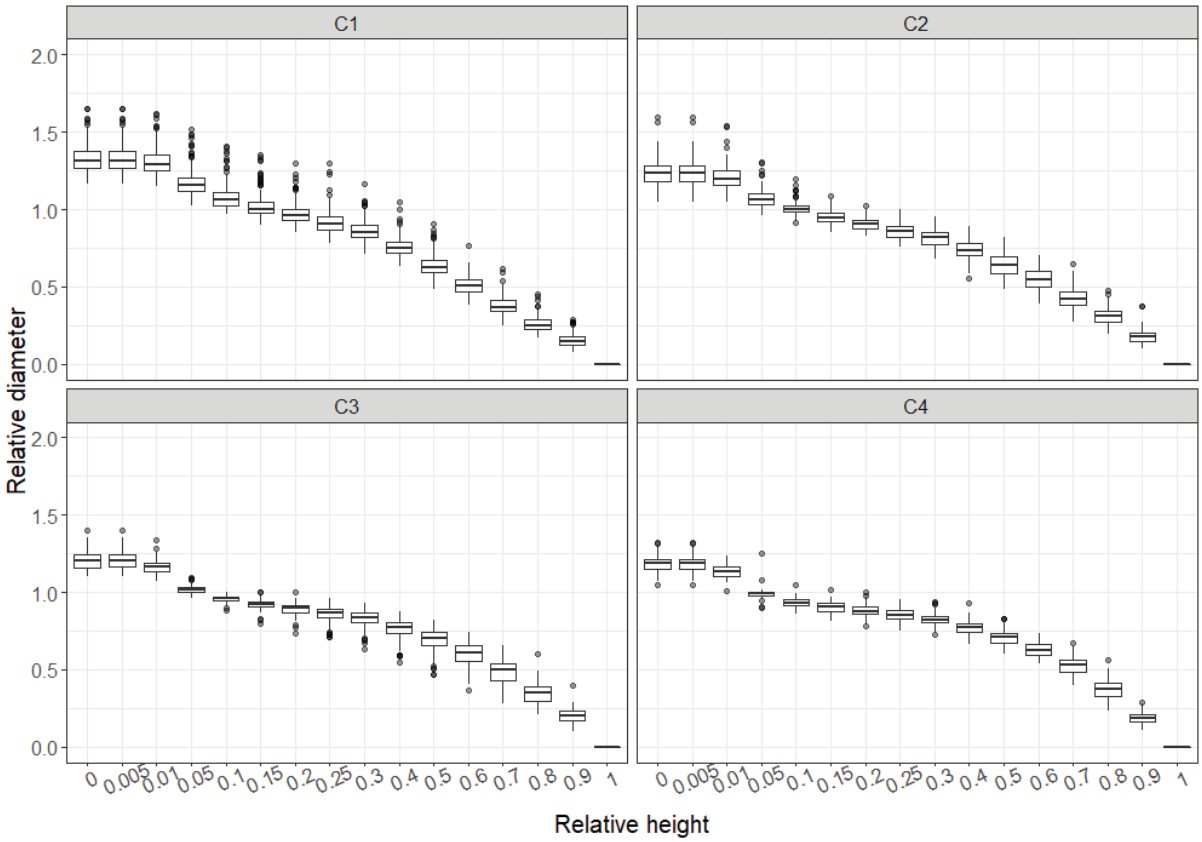
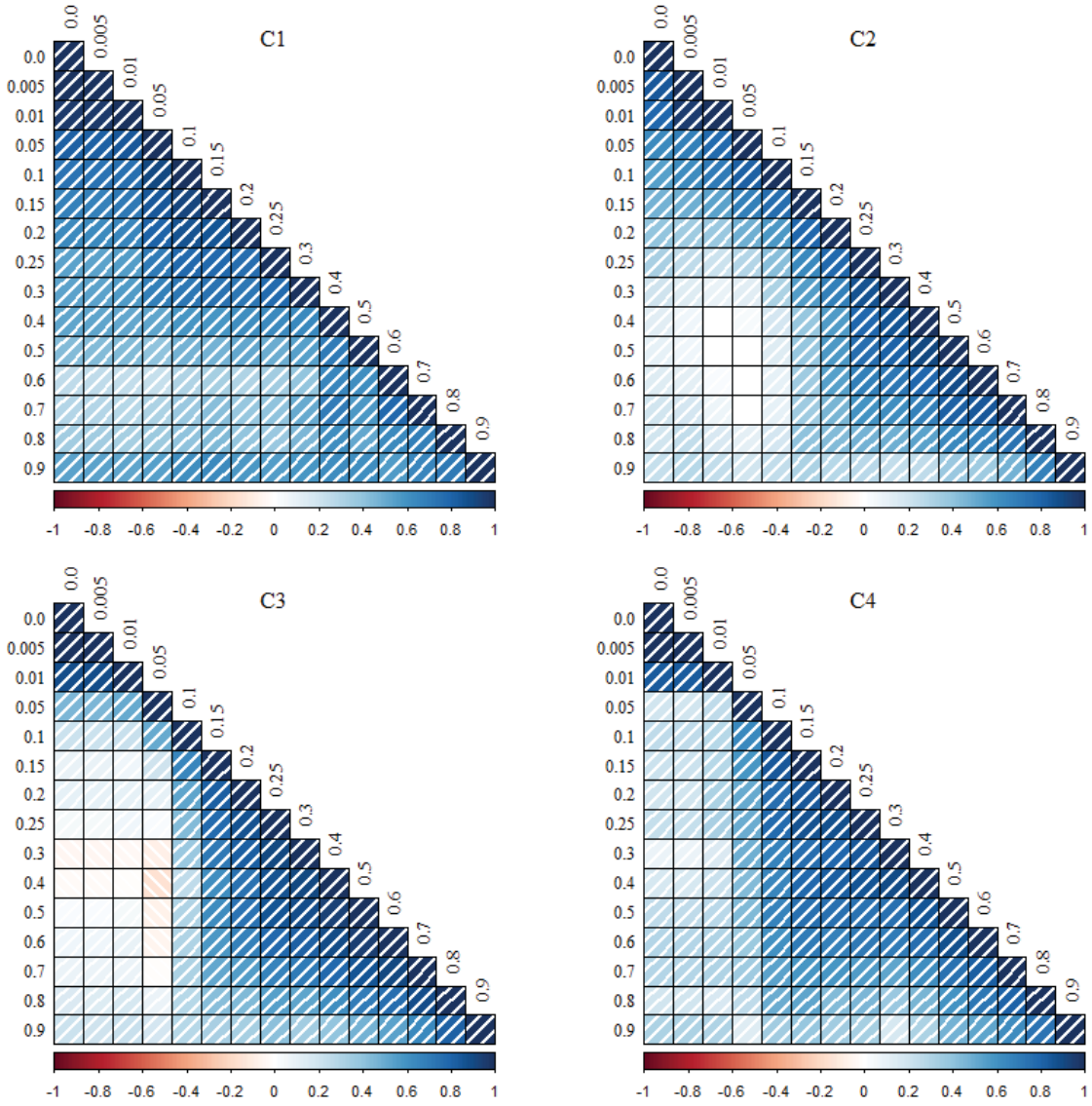


FIGURE 2.3 allows to visualize the Pearson´s linear correlation between relative diameters measured at different relative heights. The correlation intensity and pattern changed according with to the age class, but a predominant positive correlation was observed in all age classes, independent of which relative diameters we are considering. In general, a strong positive correlation was observed in C1 class, which has been smoothly decreasing as the distance between relative heights were increased, while for the others classes the changes were more abrupt. Higher correlations between relative diameters were observed for intermediate and upper portions of the stem for C2, C3 and C4 classes.

FIGURE 2.3 - CORRELOGRAM FOR REPEATED MEASURES OF RESPONSE VARIABLE RELATIVE DIAMETER BY RELATIVE HEIGHT AND AGE CLASS



2.3.2. STEM TAPER MODEL

The covariance generalized linear models (CGLM) framework is based on a second-moment assumptions. In this sub-section, we specify the non-linear predictor and a matrix linear predictor for the response relative diameter. We also present the estimated parameters for the components of the model and measures of goodness-of-fit for each modeling strategy.

A summary of the fitted non-linear predictor and a precision measure for each modeling strategy are presented in TABLE 2.1. The non-linear predictor was previously specified by the segmented non-linear model proposed by Max & Burkhart (1976). The parameter estimates were significant at 5% level, except $\hat{\beta}_1$ for *MmStr* modeling strategy.

Nevertheless, we decided to keep the fitted model, once that the non-significant parameter do not complete invalided the inferences, and we maintain the model structure. The parameter $\hat{\alpha}_1$ represented the first inflexion point of the tree stem and was similar for all strategies, ranging from 0.08 to 0.10. However, the second inflexion point $\hat{\alpha}_2$ ranged from 0.39 to 0.87, suggesting that the models can lead us to a different second inflection point. Even the parameter estimates being quite different when we compared the modeling strategies, the MSE measure was about 0.0061, suggesting a similar performance for describing an average profile stem tree.

TABLE 2.1 - PARAMETER ESTIMATES, STANDARD ERRORS (SE), Z-STATISTICS AND ROOT MEAN SQUARE ERROR (MSE) BY STRATEGIES

Parameter	Estimates	SE	Z-statistics	MSE
		VarStr		
$\hat{\beta}_1$	1.9593	0.6274	3.1232	0.00618
$\hat{\beta}_2$	-1.9762	0.3389	-5.8311	
$\hat{\beta}_3$	19.4995	1.9231	10.1395	
$\hat{\beta}_4$	1.5385	0.3284	4.6837	
$\hat{\alpha}_1$	0.1030	0.0053	19.2654	
$\hat{\alpha}_2$	0.8215	0.0278	29.5600	
		CovStr		
$\hat{\beta}_1$	3.3519	1.3324	2.5157	0.00622
$\hat{\beta}_2$	-2.7030	0.7041	-3.8392	
$\hat{\beta}_3$	23.8337	0.8543	27.8976	
$\hat{\beta}_4$	2.2442	0.6941	3.2331	
$\hat{\alpha}_1$	0.0900	0.0017	52.0427	
$\hat{\alpha}_2$	0.8726	0.0260	33.5535	
		RwStr		
$\hat{\beta}_1$	-0.1515	0.0337	-4.4919	0.00617
$\hat{\beta}_2$	-0.7929	0.0233	-33.9522	
$\hat{\beta}_3$	23.3802	1.0536	22.1907	
$\hat{\beta}_4$	1.0102	0.0902	11.2048	
$\hat{\alpha}_1$	0.0835	0.0023	35.9457	
$\hat{\alpha}_2$	0.3918	0.0163	23.9840	
		MmStr		
$\hat{\beta}_1$	-0.0444	0.0415	-1.0692	0.00617
$\hat{\beta}_2$	-0.8575	0.0284	-30.1678	
$\hat{\beta}_3$	21.9142	1.0635	20.6050	
$\hat{\beta}_4$	0.7932	0.0026	35.3354	
$\hat{\alpha}_1$	0.0902	0.0452	17.5588	
$\hat{\alpha}_2$	0.4750	0.0192	25.0412	

The components of the matrix linear predictor were selected by the score information criterion (SIC) using the stepwise procedure. The linear combination of matrices for *VarStr*

modeling strategy was composed by an identity matrix, related to the intercept of the model, and the main effect of relative height ($\mathbf{H_r}$) and its second-order effect ($\mathbf{H_r^2}$). However, the apparently interaction between the covariates height and age observed in FIGURE 2.2 was non-significant, once that the matrices of interaction effects $\mathbf{H_r : A}$ or $\mathbf{H_r^2 : A^2}$ was not selected for composing the covariance model. For *CovStr* modeling strategy, the matrix linear predictor was composed by an identity matrix combined with the Euclidean distance between pairs of observations ($\mathrm{ED_o}$) and moving average model of order 1, 2 and 3. The components for *RwStr* and *MmStr* strategies were previously defined in the subsection 2.2.4.

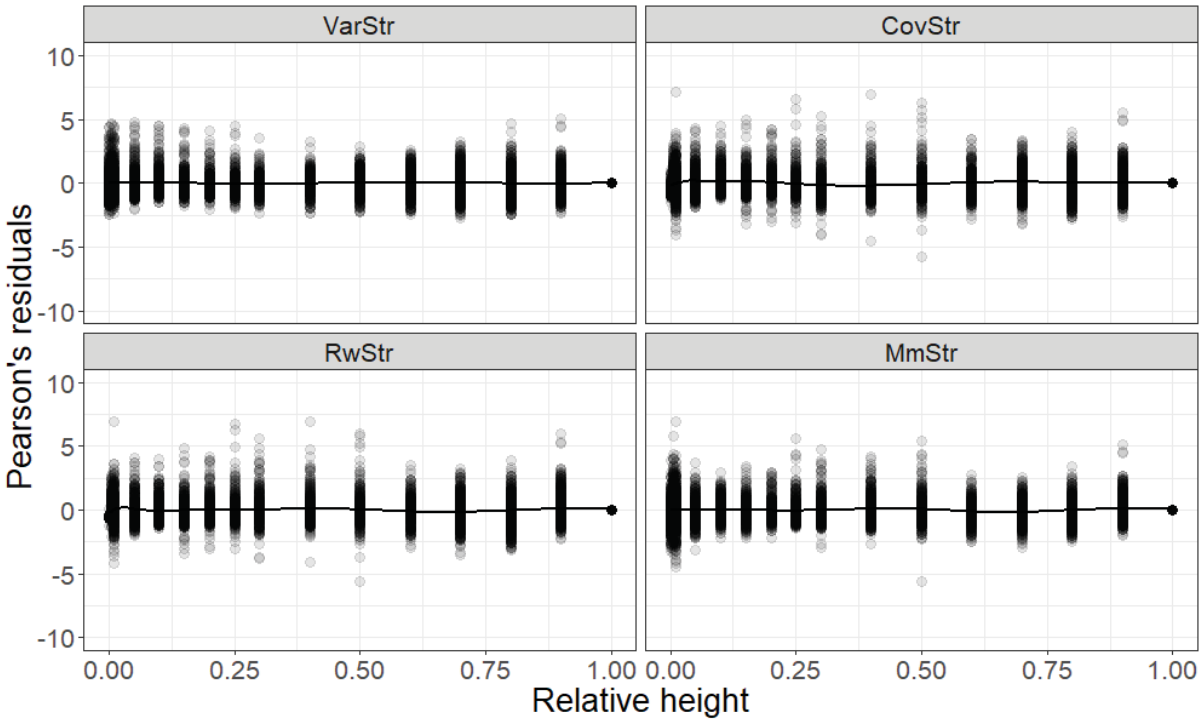
Parameter estimates for matrix linear predictor and measures of goodness-of-fit are presented in TABLE 2.2. The parameter estimates were significant at 95% confidence level for all modeling strategies. The performance of the fitted models in explaining the tree stem form was not similar when we calculated statistics based on plausibility measure. The highest value of log-likelihood (logLik), as well the lowest value for Akaike (AIC) and Bayesian (BIC) information criterion were obtained in *CovStr*, followed by *RwStr*, *MmStr* and *VarStr* strategies. These results indicated that the *CovStr* is the most suitable strategy for modeling the covariance matrix of the diameters measured at different heights. However, this fact does not invalid the application of the models. Moreover, the estimated spatial autocorrelation parameter in *RwStr* strategy was 0.9807, with confidence intervals ranged from 0.9723 to 0.9892, suggesting a high correlation between tree diameters.

TABLE 2.2 - PARAMETER ESTIMATES, STANDARD ERRORS (SE), Z-STATISTICS (Z-value), GAUSSIAN LIKELIHOOD (logLik), AKAIKE (AIC) AND BAYESIAN (BIC) INFORMATION CRITERION FOR THE MATRIX LINEAR PREDICTOR BY STRATEGIES

Parameter	Estimates	SE	Z-statistics	logLik	AIC	BIC
			VarStr			
$\hat{\tau}_0$	0.00785	0.00041	19.1058			
$\hat{\tau}_1$	0.00450	0.00155	2.9006	10281.96	-20545.92	-20484.46
$\hat{\tau}_2$	-0.01235	0.00121	-10.1782			
			CovStr			
$\hat{\tau}_0$	0.00629	0.00027	23.3944			
$\hat{\tau}_1$	0.01273	0.00063	20.2913			
$\hat{\tau}_2$	-0.00728	0.00041	-17.7324	12100.52	-24179.04	-24103.92
$\hat{\tau}_3$	-0.00164	0.00012	-13.9066			
$\hat{\tau}_4$	-0.00031	0.00004	-8.7549			
			RwStr			
$\hat{\tau}_0$	553.55150	86.80770	6.3768	11953.55	-23891.10	-23836.47
$\hat{\tau}_1$	542.89220	87.43253	6.2093			
			MmStr			
$\hat{\tau}_0$	0.00140	0.00011	12.7196			
$\hat{\tau}_1$	0.19378	0.01383	14.0073	11353.39	-22686.78	-22618.49
$\hat{\tau}_2$	0.12920	0.00927	13.9333			
$\hat{\tau}_3$	-0.15606	0.01125	-13.8734			

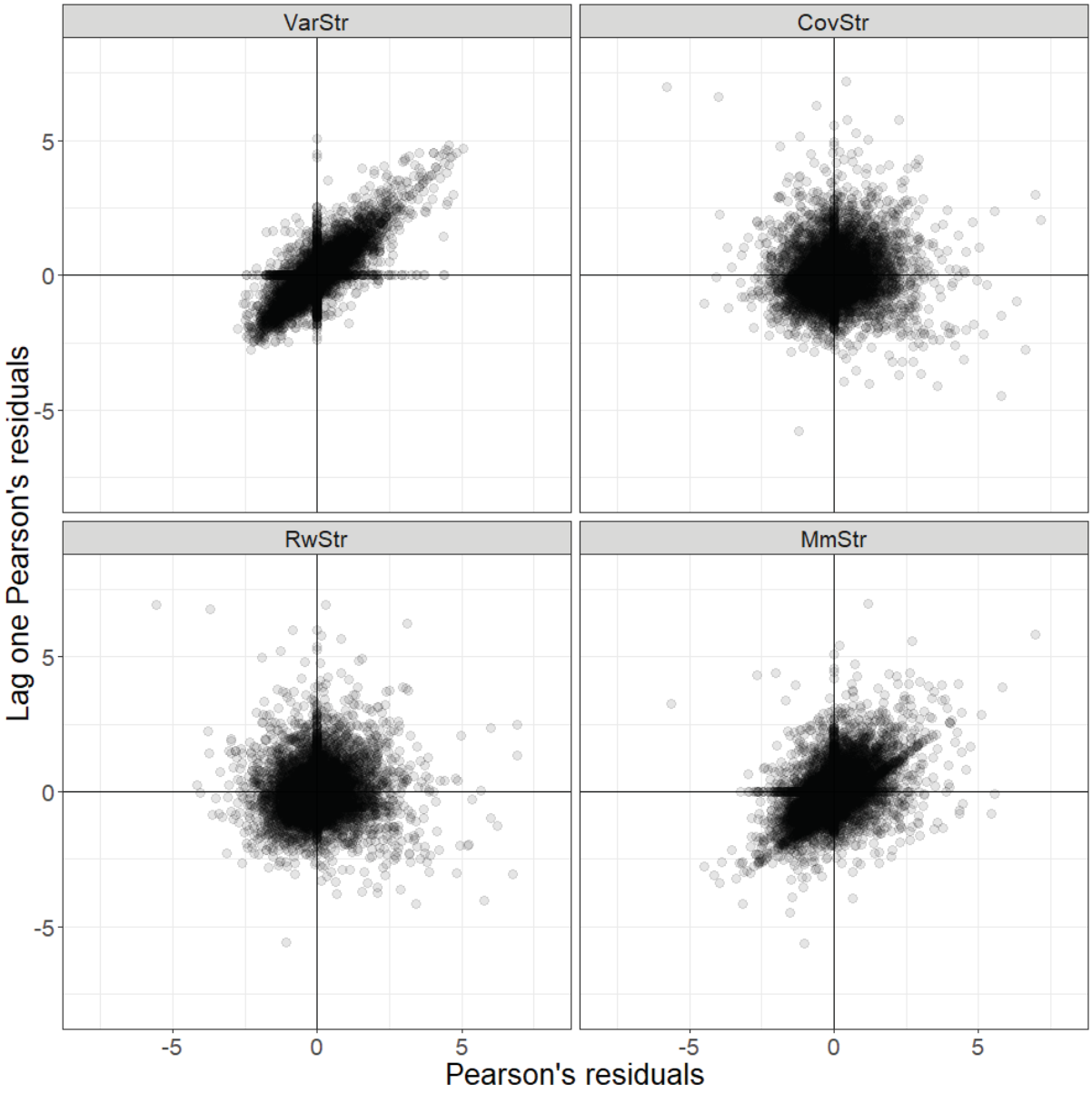
Homogeneous residual variance pattern was observed for all strategies, as given in FIGURE 2.4. Specially for *VarStr* modeling, the graphical analysis of Pearson's residuals indicated a constant pattern over the relative height, while more discrepant observations were observed for others approaches. Nevertheless, the fitted smoothed curve was constant over the relative height for all models, confirming that the models performed well.

FIGURE 2.4 - PEARSON'S RESIDUALS BY RELATIVE HEIGHT FOR DIFFERENT MODELING STRATEGIES AND FITTED SMOOTH CURVE IN SOLID LINE



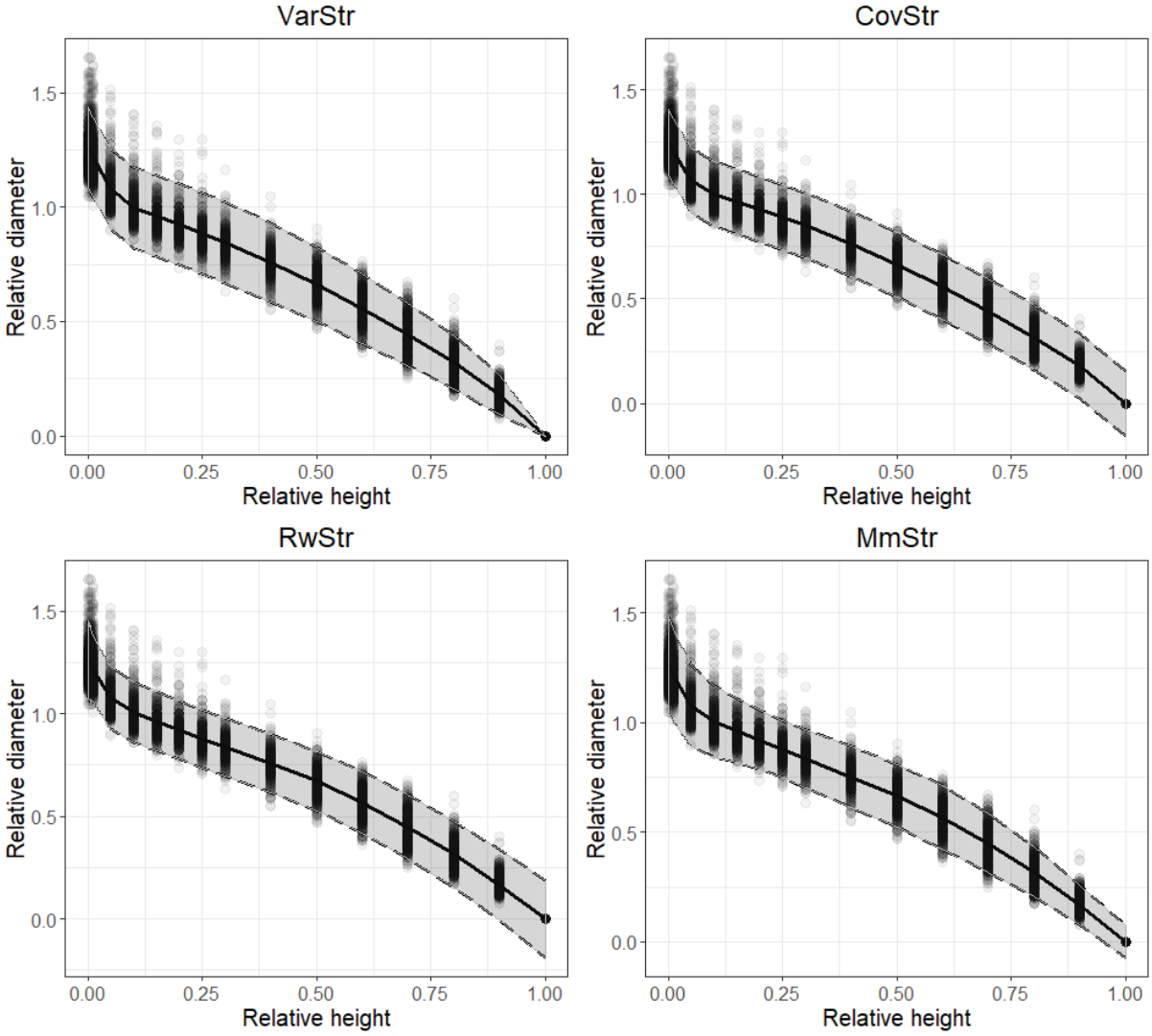
When we evaluated the correlation between residuals in FIGURE 2.5, we noticed that both *CovStr* and *RwStr* strategies were completely able to account the dependence among relative diameters and successfully removed the autocorrelation. In certain way, these results were expected, once that we defined a linear combination of matrices for explicitly deal with the dependence among observations when we specified the matrix linear predictor. However, as we did not include any dependence structure for *VarStr* strategy, and the variance was directly modeled by covariates, an autocorrelation pattern was still observed. Similar pattern was obtained for *MmStr* approach, but in a moderate intensity.

FIGURE 2.5 - CORRELATION BETWEEN LAG ONE PEARSON'S RESIDUALS FOR RESPONSE VARIABLE RELATIVE DIAMETER FITTED FOR DIFFERENT MODELING STRATEGIES



The predicted mean stem profile is showed in FIGURE 2.6. Even the modeling strategies showing different estimates for the same parameters of the non-linear predictor, the predicted relative diameter was similar for all models. These results are according to the values of the MSE previously obtained. The uncertainty expressed by 95% confidence intervals for response were quite different among strategies, being a directly effect from the choice of the matrix linear predictor. However, due to the symmetric confidence interval around the predictions, negative lower bound were obtained for all modeling strategies in the top of the tree.

FIGURE 2.6 - UNCERTAINTY IN THE PREDICTIONS. OBSERVED VALUES (FULL CIRCLES), FITTED VALUES (SOLID LINES) AND 95% CONFIDENCE INTERVALS (DASHED LINES) FOR RESPONSE VARIABLE RELATIVE DIAMETER FOR DIFFERENT MODELING STRATEGIES

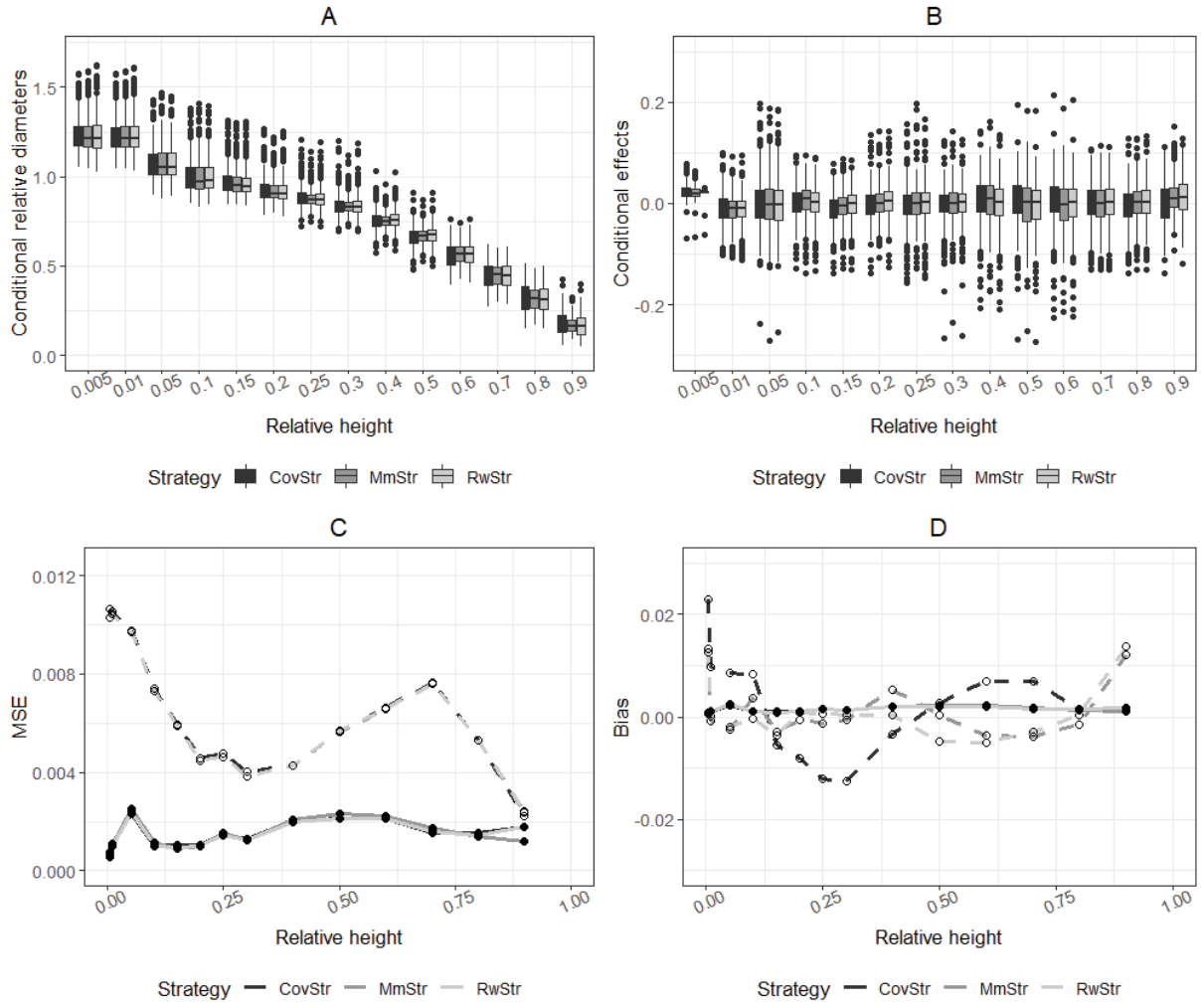


2.3.3. MARGINAL AND CONDITIONAL PREDICTIONS

FIGURE 2.7 shows conditional relative diameters (A) from a marginal model fitting and the conditional effects computed from equation (5) by relative height (B). The modeling strategies presented similar behavior for both analyses. An overestimate of the conditional effects for relative diameters until 40% of total height and for the top of the tree was observed for all modeling strategies, due to the asymmetries on the boxplots. Nevertheless, even no assumptions is required in our approach, Shapiro-Wilk test indicated the conditional effects has a normal distribution by relative height, for significance level of 5%.

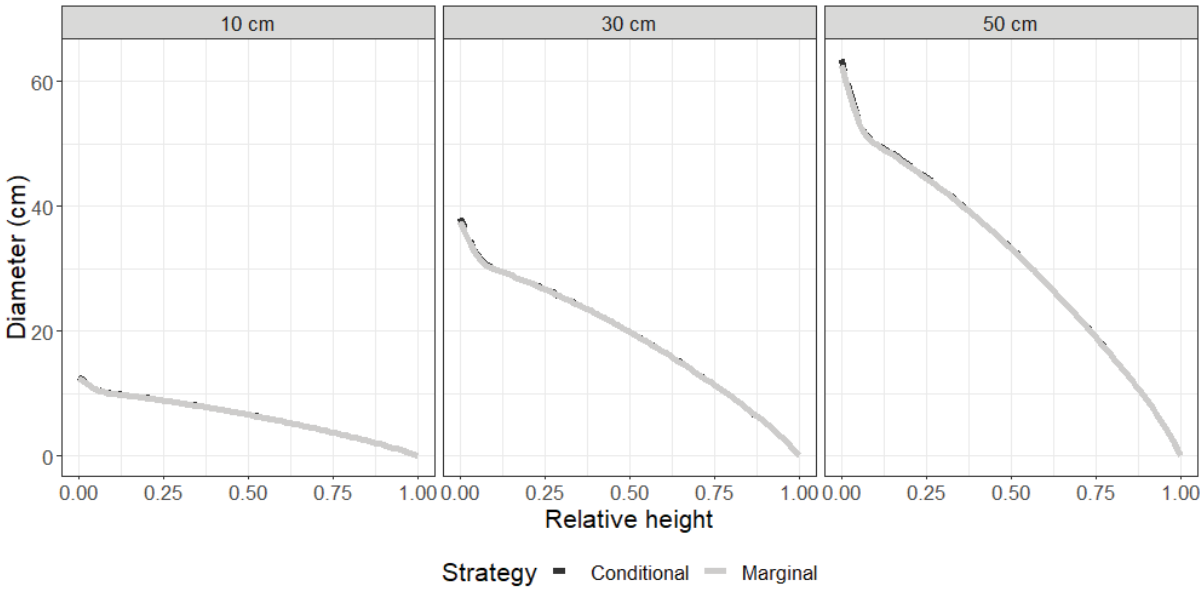
Mean squared error (C) and Bias (D) for predictions were also presented in FIGURE 2.7. As we expected, both error measures suggested that the diameter predictions from conditional model specification provides better performance when compared it to the marginal specification. Conditional effects were also able to stabilize the errors, being almost constant over the stem. In general, higher errors were observed between 30 to 70% of relative height. The analysis was not performed for *VarStr*, once this strategy does not present correlation parameters.

FIGURE 2.7 - PREDICTED CONDITIONAL RELATIVE DIAMETERS (A) AND PREDICTED CONDITIONAL EFFECTS (B) FOR DIFFERENT MODELING STRATEGIES. MEAN SQUARED ERROR (C) AND BIAS (D) FOR MARGINAL (DASHED LINES WITH EMPTY CIRCLES) AND CONDITIONAL (SOLID LINES WITH FULL CIRCLES) PREDICTIONS OF RESPONSE VARIABLE RELATIVE DIAMETER FOR DIFFERENT MODELING STRATEGIES



We selected the *CovStr* strategy as the best fitted model based on the previously results. Conditional and marginal predictions for relative diameter were also estimated for new individuals. The predictions were performed for trees with 10 cm, 30 cm and 50 cm of diameter, but conditioned only to the diameter at breast height. The predicted values were quite similar between approaches, with more relevant differences until 25% of relative height, as shown in FIGURE 2.8. This means that the diameter at breast height are not highly correlated with diameters measured after 25% of the total height.

FIGURE 2.8 - CONDITIONAL AND MARGINAL PREDICTIONS OF RESPONSE VARIABLE RELATIVE DIAMETER FOR *CovStr* MODELING STRATEGIES CONSIDERING DIAMETER AT BREAST HEIGHT OF 10, 30 AND 50 CM



2.4. DISCUSSION

Multiple diameters are measured at regular or irregular distances along tree stem for fitting taper functions. Thus, it is reasonable to expect some autocorrelation between diameters due to the multiple measurements taken from the same sample tree (KUBLIN et al., 2013). Taper functions are frequently composed by polynomial terms and other transformations of the same predictor variable, being the relative height the most common covariate. However, this procedure can cause high multicollinearity among independent variables (KOZAK, 1997). In this context, the fundamental assumptions of independent observations, errors with normal distribution and constant variance are usually violated in linear and non-linear regression

models. From the statistical point of view, the parameter estimates still remaining unbiased and consistent, but the minimum variance property it is not reached, that implies in unreliable hypothesis tests and inferences. In this research, our main focus was present a new taper function modeling approach based on covariance generalized linear models (CGLM). We developed marginal linear covariance models for handling explicitly with non-constant residual variance and autocorrelation patterns by including a matrix linear predictor on the model specification.

Parameter estimates in the non-linear predictor changed according to the modeling strategies. In general, standard errors associated to *RwStr* were smaller, while larger values were obtained in *CovStr*. The covariance matrix has special importance for the CGLM approach. While this structure has a little influence on the mean parameter estimators, the associated standard error directly depends on the correct choice of the covariance structure (BONAT & JØRGENSEN, 2016) for a reliable confidence interval and hypothesis tests. This feature of our modeling approach explains the similar behavior of the marginal stem profile among different strategies.

Then, whether the segmented polynomial model is changed on the non-linear predictor, it is natural to expect a different behavior in diameter predictions. As alternative to the traditional models for mean response, flexible semi-parametric taper models have been explored to describe the stem profiles and volume predictions (KUBLIN et al., 2013). The CGLM framework also allow to incorporate splines and penalized spline in the mean structure (BONAT et al., 2017), what is an interesting topic to be explored in future researches.

The *MmStr* strategy was formulated to take into account repeated measures effects within-tree in a mixed model design. In previous data analysis, our marginal specification of the segmented non-linear model was compared to a conditional specification based on mixed-effect model. The log-likelihood was equal for both approaches, indicating an equivalence between the modeling strategies. However, the small differences in the parameter estimates were related to the different fitting algorithms used by each statistical framework. In previous research, Bonat (2018) studied the efficiency property of a Gaussian linear mixed-effects model and its marginal specification fitted in the mcglm package. The author reported that the log-likelihood was equal in both cases and provided virtually the same estimates for the regression and dispersion parameters. However, due to the robust specification of the MCGLM framework, and a not fully efficient estimating function for the estimation of the dispersion parameters, standard errors associated with the dispersion parameters were larger.

The *VarStr* strategy allowed to model the variability of the relative diameters using the covariates tree relative height and age. However, just the diagonal matrices representing the positions where the diameter was taken ($\mathbf{H_r}$ and $\mathbf{H_r^2}$) captured the non-constant residual patterns, indicating a second-order polynomial relationship between the mean response and the variance of the relative diameters. In order to investigate the non-significant effect of age, we performed a graphical analysis about the variance over relative heights and performed 95% confidence intervals by age class, but not shown in this paper. Independent of age class, confidence intervals contain almost all the values of variance. Thus, even the variance having an apparent differentiated behavior when we changed the age, it was not statistically significant.

Our study showed that the *CovStr* strategy performed well for tree stem taper modeling. The components of the matrix linear predictor based on Euclidean distance between observations and moving average model of order 1, 2, and 3 provided most explanation about the autocorrelation of the relative diameters. Euclidean distance matrix has special importance because the relative heights were not equally spaced over the stem. FIGURE 2.3 also suggested different components for the matrix linear predictor whether fitting the models by age class, once that the positive correlation pattern is changing over time. When we performed a correlogram matrix for entire data set e not just for age class, not shown in this paper, a negative correlation between diameters taken in the relative heights among 0-10% and 60-80% was observed. Similar pattern was also reported by Diéguez-Aranda et al. (2006) for Scot Pine in northwestern Spain. These authors suggested the autocorrelation within-tree may be also explained by effects of forest stands conditions, particularly stand density.

A very interesting result on the parameter estimates of the matrix linear predictor were found. As we mentioned, the CGLM are based on a marginal specification, i.e., they are a class of models that require a mean and covariance components specification. In this case, we did not test the null hypothesis of the dispersion parameters $\tau = 0$ on the boundary of the parameter space, as did by commonly likelihood ratio, Wald and score test (BONAT, 2017). In this case, it is natural to obtain values for the dispersion parameters smaller than zero, caused by sample variation. Fiorentin et al. (2020) reported similar results on the dispersion parameters for a jointly modeling of height and volume for *Araucaria angustifolia*.

Due to the features of each modeling strategy, we combined the components of the matrix linear predictor selected for *VarStr* and *CovStr* strategies. Our initial expectation was simultaneously handling with autocorrelation and non-constant variance pattern in order to obtain a more complete model using a general formulation. However, the main effect of relative height ($\mathrm{H_r}$) and its second-order effect ($\mathrm{H_r^2}$) were jointly non-significant in the combined model,

resulting in the *CovStr* strategy previously formulated. This fact suggested that it is not necessary a variance structure whether the correlation patterns is correctly specified in the regression model. However, this result can be also affected by the test hypothesis that assume a normal distribution for the variance components. Bonat & Jørgensen (2016) reported that the misspecification of the covariance structures can conduct to an underestimation and overestimation of the standard errors associated with the regression parameters for the linear predictor in the CGLM framework.

A constant standard error of prediction over the stem was observed for *CovStr* modeling strategy. However, the range of standard error showed a quadratic pattern with different concavity for *RwStr* and *MmStr*, while *RwStr* presented a cubic trend, what is directly related to the selected components of the matrix linear predictor (TABLE 2.2). Thus, these results mean that the choice of covariance matrix directly influences the uncertainty analysis related to the diameter predictions over the tree stem. The uncertainty analysis also showed that the lower limit of the confidence intervals for response were negative for diameters at the top of the trees. For overcome these limitations, a link function component can be included in the regression model as suggested by Fiorentin et al. (2020). The authors applied a multivariate CGLM approach for jointly marginal modeling of the diameter-height relationship and individual volume, where identity and logarithm link functions were suitable, respectively.

The methodology developed for predicting the conditional diameter effects for new individuals are quite easy to be applied in practical analysis, being also recommended for non-forestry data. After fitting the specified model, as the *CovStr* strategy, a grid of relative height for diameter predictions can be defined and a general covariance matrix are easily built by using the parameter estimates. Thus, the advantage of our model is to predict the diameter for any relative height conditionate to the diameter at breast height. The conditional effects generated from a marginal specification increased the predictions from the data set. However, we did not observe relevant differences from conditional and marginal predictions for new individuals. In this sense, additional diameter measures should be taken over the stem height for generating the conditional effects, in a similar idea of model calibration, quite common in the mixed-effects models (see CASTEDO-DORADO et al., 2006; LEJEUNE et al., 2009). The disadvantage of collecting supplementary diameters is increasing the costs for measuring and the time for processing the data.

2.5. CONCLUSION

We presented a statistical approach for modeling tree stem taper based on the covariance generalized linear models. Our approach allowed a flexible modeling of covariance, which is an important part for developing suitable taper function, once that we can choose a large set of covariance structures with different correlation patterns. Besides, the common non-constant residual variance of the taper models was directly modeled by covariates.

The advantage of our marginal specification of the tree stem taper model is the directly interpretation of covariates effects on the population mean for both regression and dispersion parameters. In addition to the new methodology for stem taper modeling, another advantage of our approach is obtaining a robust taper model, which can be applied with high precision to a large variety of forest stands conditions.

The main advantages of the *CovStr* approach is the easy formulation of the model. Once that we select the linear predictor, the components of matrix linear predictor are selected by score information criterion from a set of covariance structures. The results suggest that the Euclidean distance matrix is a fundamental component, especially when the relative heights are not equally spaced over the stem. The moving average structure of order one to three indicate that the correlation among diameters decrease over the stem.

Conditional predictions from the marginal model specification improve the predictions of response variable relative diameter. Besides, the conditional prediction for new individuals can be easily generated by using covariance generalized linear models for any relative height. We also recommend to include more than one diameter measures over the stem when conditionate the model. However, additional measures can introduce higher costs to the forest inventories and higher time demand for collecting the data in the field.

The uncertainty in diameter estimation are easily quantify in covariance generalized linear models by the confidence intervals. This procedure is important to ensure a suitable forestry management planning. We recommend including a non-parametric bootstrap approach as an additional tool for complementing the uncertainty analysis.

Future topics for research include to extend the analysis of stem taper data using covariance generalized linear model for merchantable volume prediction. Besides, to adapt the modeling framework presented to model individual tree growth with univariate and multivariate response variables. An interesting topic to be researched is to investigate the biomass dynamic considering a spatial, temporal or spatial-temporal model with different dependence structures. The multivariate case of covariance generalized linear model has great potential to be applied

to height-diameter and volume modeling, where link functions and variance functions can be easily incorporated in the biometric models.

REFERENCES

ARIAS-RODIL, M.; CASTEDO-DORADO, F.; CÁMARA-OBREGÓN, A.; DIÉGUEZ-ARANDA, U. Fitting and calibrating a multilevel mixed-effects stem taper model for Maritime Pine in NW Spain. **PlosOne**, v. 10, 2015a.

ARIAS-RODIL, M.; DIÉGUEZ-ARANDA, U.; PUERTA, F.R.; LÓPEZ-SÁNCHEZ, C.A.; LÍBANO, E.C.; OBREGÓN, A.C.; CASTEDO-DORADO. Modeling and localizing a stem taper function for *Pinus radiata* in Spain. **Canadian Journal of Forest Research** , v. 45, p. 647–658, 2015b.

ARIAS-RODIL, M.; DIÉGUEZ-ARANDA, U.; BURKHART, H.E. Effects of measurement error in total tree height and upper-stem diameter on stem volume prediction. **Forest Science**, v. 63, n. 3, p. 250–260, 2017.

BERGER, A.; GSCHWANTNER, T.; MACROBERTS, R.E.; SCHADAUER K. Effects of measurement errors on individual tree stem volume estimates for the Austrian National Forest Inventory. **Forest Science**, v. 60, n. 1, p. 14–24, 2014.

BONAT, W.H. Modelling mixed types of outcomes in additive genetic models. **The international journal of biostatistics**, v. 13, n. 2, p. 1–16, 2017.

BONAT, W.H. Multiple response variables regression models in R: The mcglm Package. **Journal of Statistical Software**, v. 84, n. 4, 2018.

BONAT, W.H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society. Series C: Applied Statistics**, v. 65, n. 5, p. 649–675, 2016.

BONAT, W.H.; OLIVERO, J.; GRANDE-VEJA, M.; FARFÁN A.; FA, J.E. Modelling the Covariance Structure in Marginal Multivariate Count Models: Hunting in Bioko Island. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 22, n. 4, p. 446–464, 2017.

BOSE, A.K.; WEISKITTEL, A.; KUEHNE, C.; WAGNER, R.G.; TURNBLOM, E.; BURKHART, H.E. Tree-level growth and survival following commercial thinning of four major softwood species in North America. **Forest Ecology and Management**, v. 427, p. 355–364, 2018.

BURKHART, H.E.; TOMÉ, M. **Modeling forest trees and stands**. New York: Springer, 2012. 457 p.

CAO, Q.V.; WANG, J. Evaluation of methods for calibrating a tree taper equation. **Forest Science**, v. 61, n. 2, p. 213–219, 2015.

CASTEDO-DORADO, F.; DIÉGUEZ-ARANDA, U.; ANTA, M.B.; RODRÍGUEZ, M.S.; VON GADOW, K. A generalized height-diameter model including random components for radiata pine plantations in northwestern Spain. **Forest Ecology and Management**, v. 229, n. 3, p. 202–213, 2006.

DIÉGUEZ-ARANDA, U., CASTEDO-DORADO, F.; ÁLVAREZ-GONZÁLEZ, J.G.; ROJO,

A. Compatible taper function for Scots pine plantations in northwestern Spain. **Canadian Journal of Forest Research**, v. 36, n. 5, p. 1190–1205, 2006.

FIORENTIN, L.D.; BONAT, W.H.; PELISSARI, A.L.; MACHADO, S.A.; TÉO, S.J. Modelagem marginal conjunta da altura e volume para *Araucaria angustifolia*. **Biofix Scientific Journal**, v. 5, n. 1, p. 121–129, 2020.

FORTIN, M.; ROBERT, N.; MANSO, R. Uncertainty assessment of large-scale forest growth predictions based on a transition-matrix model in Catalonia. **Annals of Forest Science**, v. 73, n. 4, p. 871–883, 2016.

FORTIN, M.; SCHNEIDER, R.; SAUCIER, J. Volume and error variance estimation using integrated stem taper models. **Forest Science**, v. 59, n. 3, 2013.

GOMAT, H.Y.; DELEPORTE, P.; MOUKINI, R.; MIALOUNGUILA, G.; OGNOUABI, N.; SAYA, A.R.; VIGNERON, P.; SAINT-ANDRE, L. What factors influence the stem taper of Eucalyptus: Growth, environmental conditions, or genetics? **Annals of Forest Science**, v. 68, n. 1, p. 109–120, 2011.

GÓMEZ-GARCÍA, E.; CRECENTE-CAMPO, F.; DIÉGUEZ-ARANDA, U. Selection of mixed-effects parameters in a variable-exponent taper equation for birch trees in northwestern Spain. **Annals of Forest Science**, v. 70, n. 7, p. 707–715, 2013.

KOZAK, A. Effects of multicollinearity and autocorrelation on the variable-exponent taper functions. **Canadian Journal of Forest Research**, v. 27, n. 5, p. 619–529, 1997.

KOZAK, A. My last words on taper equations. **Forestry Chronicle**, v. 80, n. 4, p. 507–515, 2004.

KUBLIN, E.; BREIDENBACH, J.; KÄNDLER, G.A. flexible stem taper and volume prediction method based on mixed-effects B-spline regression. **European Journal of Forest Research**, v. 132, n. 5–6, p. 983–997, 2013.

LEJEUNE, G., UNG, C.H.; FORTIN, M.; GUO, X.J.; LAMBERT, M.C.; RUEL, J.C. A simple stem taper model with mixed effects for boreal black spruce. **European Journal of Forest Research**, v. 128, n. 5, p. 505–513, 2009.

LEE, Y.; NELDER, J.A. Conditional and marginal models: Another view. **Statistical Science**, v. 19, n. 2, p. 219–238, 2004.

LI, R.; WEISKITTEL, A. Development and evaluation of regional taper and volume equations for the primary conifer species in the Acadian Region of North America. **Annals of Forest Science**, v. 67, p. 21–24, 2010.

MACFARLANE, D.W.; WEISKITTEL, A.R. A new method for capturing stem taper variation for trees of diverse morphological types. **Canadian Journal of Forest Research**, v. 46, n. 6, p. 804–815, 2016.

MÄKINEN, H.; JYSKE, T.; NÖJD, P. Dynamics of diameter and height increment of Norway spruce and Scots pine in southern Finland. **Annals of Forest Science**, v. 75, n. 1, p. 1–11, 2018.

MACPHEE, C.; KERSHAW, J.A.; WEISKITTEL, A.R.; GOLDING, J.; LAVIGNE, M.B. Comparison of approaches for estimating individual tree height-diameter relationships in the Acadian forest region. **Forestry: An international Journal of Forest Research**, v. 91, n. 1, p. 132–146, 2018.

MANSO, R.; NINGRE, F.; FORTIN, M. Simultaneous prediction of plot-level and tree-level harvest occurrences with correlated random effects. **Forest Science**, v. 64, n. 5, p. 461–470, 2018.

MAX, T.; BURKHART, H. Segmented polynomial regression applied to taper equations. **Forest Science**, v. 22, n. 3, p. 283–289, 1976.

MCROBERTS, R.E.; WESTFALL, J,A. Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. **Forest Science**, v. 60, n. 1, p. 34–42, 2014.

NASCIMENTO, R.G.M.; MACHADO, S.A.; FIGUEIREDO FILHO, A.; HIGUCHI, N. A growth and yield projection system for a tropical rainforest in the Central Amazon, Brazil. **Forest Ecology and Management**, v. 327, p. 201–208, 2014.

OIJEN, V.M. Bayesian methods for quantifying and reducing uncertainty and error in forest models. **Current Forestry Reports**, v. 3, n. 4, p. 269–280, 2017.

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2019.

RIOFRÍO, J.; DEL RÍO, M.; BRAVO, F. Mixing effects on growth efficiency in mixed pine forests. **Forestry: An international Journal of Forest Research**, v. 90, n. 3, p. 381–392, 2017.

SABATIA, C.O. Use of upper stem diameters in a polynomial taper equation for New Zealand radiata pine: an evaluation. **New Zealand Journal of Forestry Science**, v. 46, n. 1, 2016.

SABATIA, C.O.; BURKHART, H.E. On the Use of Upper Stem Diameters to Localize a Segmented Taper Equation to New Trees. **Forest Science**, v. 61, n. 3, 2015.

SEKI, M.; SAKICI, O.E. Dominant height growth and dynamic site index models for crimean pine in the Kastamonu-Taşköprü region of Turkey. **Canadian Journal of Forest Research**, v. 47, n. 11, p. 1441–1449, 2017.

SHARMA, M.; REID, D.E.B. Stand height/site index equations for jack pine and black spruce trees grown in natural stands. **Forest Science**, v. 64, n. February, p. 33–40, 2017.

SHARMA, R. P.; VACEK, Z.; VACEK, S.; JANSA, V.; KUCERA, M. Modelling individual tree diameter growth for Norway spruce in the Czech Republic using a generalized algebraic difference approach. **Journal of Forest Science**, v. 63, n. 5, p. 227–238, 2017.

STOKLOSA, J.; GIBB, H.; WARTON, D.I. Fast forward selection for generalized estimating equations with a large number of predictor variables. **Biometrics**, v. 70, n. 1, p. 110–120, 2014.

TENZIN, J.; TENZIN, K.; HASENAUER, H. Individual tree basal area increment models for broadleaved forests in Bhutan. **Forestry: An international Journal of Forest Research**, v. 90, n. 3, p. 367–380, 2017.

TRINCADO, G.; VANDERSCHAAF, C.L.; BURKHART, H.E. Regional mixed-effects height-diameter models for loblolly pine (*Pinus taeda* L.) plantations. **European Journal of Forest Research**, v. 126, n. 2, p. 253–262, 2007.

VERBEKE, G.; FIEUWS, S.; MOLENBERGHS, G. The analysis of multivariate longitudinal data: A review. **Stat methods Med Res**, v. 23, n. 1, p. 42–59, 2014.

WESTFALL, J.A.; SCOTT, C.T. Taper models for commercial tree species in the northeastern United States. **Forest Science**, v. 56, n. 6, p. 515–528, 2010.

WESTFALL, J.A.; MCROBERTS, R.E.; RADTKE, P.J.; WEISKITTEL, A.R. Effects of uncertainty in upper-stem diameter information on tree volume estimates. **European Journal of Forest Research**, v. 135, n. 5, p. 937–947, 2016.

YANG, Y.; HUANG, S.; MENG, S.X. Development of a tree-specific stem profile model for white spruce: A nonlinear mixed model approach with a generalized covariance structure. **Forestry: An international Journal of Forest Research**, v. 82, n. 5, p. 541–555, 2009.

3. MODELAGEM MARGINAL CONJUNTA DE VOLUME E ALTURA PARA Araucaria angustifolia

RESUMO

Variáveis mensuradas em florestas normalmente apresentam algum grau de correlação. Logo, ajustar modelos para estimar variáveis biométricas de forma independente não é a abordagem mais adequada. Assim, modelos multivariados ganham relevância devido à capacidade de quantificar associações entre variáveis respostas. Nesse contexto, o objetivo da presente pesquisa foi ajustar modelos lineares generalizados de covariância multivariada (MCGLMs) no caso univariado e multivariado para estimar altura e volume de árvores. As variáveis altura (H), volume (V) e diâmetro (D) foram coletadas da *Araucaria angustifolia*. em floresta nativa, localizada no estado de Santa Catarina, Brasil. Os MCGLMs foram ajustados para estimar H e V, em abordagem univariada e multivariada. O preditor linear dos modelos foi fixado previamente em função da covariável D, para ambas as variáveis. Devido a um aparente padrão de variância não constante das duas respostas, diferentes estruturas do preditor linear de matriz foram testadas, com efeito da covariável D variando até um polinômio de grau três. Ainda, um parâmetro de potência foi estimado nas duas abordagens, com a finalidade de obter uma função de variância para cada variável. Os parâmetros estimados nas abordagens univariadas e multivariadas foram similares. Em geral, o erro padrão dos parâmetros foi menor para os modelos multivariados, sendo consequência da correlação entre as variáveis respostas. Os resultados também sugeriram que uma função de variância Poisson-Gama composta é adequada para variável V, bem como uma função constante para variável H. O modelo mais adequado foi obtido com preditor linear matricial somente em função de um parâmetro de dispersão associado a uma matriz identidade.

Palavras-chave: Distribuição Tweedie. Floresta Ombrófila Mista. Regressão multivariada.

JOINT MARGINAL MODELING OF HEIGHT AND VOLUME FOR *Araucaria angustifolia*

ABSTRACT

Variables measured in a forest usually present correlation between them. Fitting models for estimating biometric variables in an independent way is not the most suitable approach. Thus, multivariate models become interesting due to the ability of quantifying associations between response variables. In this context, the main objective of this research was to fit univariate and multivariate regression models based on multivariate covariance generalized linear models (MCGLM) for estimating the trees height and volume. The variables height (H), volume (V) and diameter (D) were obtained from *Araucaria angustifolia*, in native forest, located at Santa Catarina state, Brazil. The MCGLM were fitted for estimating H and V in univariate and multivariate approach. The linear predictor of the models was previously fixed as a function of covariate D for both responses. Due to the apparently non-constant covariance pattern for both variables, we tested different structures for the matrix linear predictor, where the effect of covariate D changing until a third-degree polynomial model. Still, a power parameter was estimated in both approaches where the aim was to obtain a variance function for each covariate. The estimated parameters of the univariate and multivariate approaches were similar for some models. In general, the standard error of the parameters was lower for multivariate models, what is a consequence of the correlation between responses variables. The results also suggested that a composed Poisson-Gama variance function is suitable for V and a constant function is required for H. The most suitable model was obtained with matrix linear predictor as a function of a dispersion parameter associated to an identity matrix.

Keywords: Tweedie distribution. Mixed Ombrophilous Forest. Multivariate regression.

3.1. INTRODUCTION

Araucaria angustifolia (Bert.) O. Ktze. is a native specie from Brazil and belongs to Araucariaceae family, being the unique representative specie of this family in Brazilian flora (MARCHIORI, 2005). *Araucaria angustifolia* occurs in diversified associations, which compromise cluster with their own characteristics, forming distinct successional stages (WEBER et al., 2017). Despite the great importance of this species for the native forests, especially that one in Southern Brazil, it is currently threatened of extinction due to the over exploration without an appropriate replacement (SCHEEREN et al., 1999).

Due to the higher economic values of *Araucaria angustifolia* timber, it is fundamental to quantify the wood stocks in forestry ecosystems. The volume is one of the most important information for evaluating the potential of a forest, since tree individual volume provides subsidies for the assessment of wood stock; and for analyzing the productive forest potential (THOMAS et al., 2006). Still, variables such as diameter at breast height and total height are fundamental attributes at tree level for many aspects of the forest management. However, tree volume and height quantification in native forest is a costly activity and time demand for collecting the data, while the diameter mensuration is a relatively simple procedure with lower costs. In this context, it is common to fit statistical models for describing the behavior of variables that are hard to obtain in the field as a function of variables, such as the tree diameter.

The most common approach for characterizing the tree height and volume in native forests or forest stands is to apply generalized linear models (FU et al., 2017), non-linear models (LAM et al., 2017) and mixed-effects models (MEHTÄTALO et al., 2015). Thus, the modeling process is usually performed in individual way, where the response variables are considered as non-correlated variables, i.e., we assume that they are independent. However, it is expected that the variables measured on the same individual present some correlation degree because the trees are biological organisms. Therefore, modeling tree height and volume independently is not the most appropriate approach.

The multivariate regression models are generalizations of the univariate models and allows to study more than two response variables simultaneously. However, applications of this class of models is quite restrict in forest research (see LAPPI, 2006). Recently, BONAT & JØRGENSEN (2016) develop the so-called multivariate covariance generalized linear models (MCGLM). The MCGLM is a class of multivariate statistical models that allow to model response variables from distinct nature simultaneously. Thereby, within the MCGLM it is possible to describe the behavior of continuous data, such as diameter at breast height, total

height and individual volume; as well as discrete data, usually obtained from count data, as the number of trees attacked by pest, and the abundance of forest species.

The main advantage of the MCGLM it is your flexibility in modeling a set of response variables from different natures simultaneously, besides quantifying the association among them by using correlation parameters. Another advantage is related to the covariance matrix, which is specified on the matrix linear predictor. This structure is quite flexible and allows to model temporal and spatial correlated data by a linear combination of known matrices, which can describe many behaviors and kind of associations. Thus, a set of correlations structures among observations can be included, besides variance functions for different type of response variables, and to model in a suitable way the natural data variability (BONAT & JØRGENSEN, 2016; BONAT et al., 2017; BONAT, 2018).

In this study, univariate and multivariate models were fitted for response variables tree total height and individual volume as a function of the covariate diameter at breast height. Still, we tested some variance structures, as well as the inclusion of variance functions. The research hypothesis is that the correlation between response variables influence the estimates and inference of the multivariate regression models. Therefore, the main goal of this paper was to analyze the fitting of univariate and multivariate regression models for describing the behavior of height and volume of *Araucaria angustifolia*, in native forest.

In the material and methods section are described the data set used as motivation and a brief introduction about the MCGLM. In that follows, we highlighted the main results and discussions about the fitted univariate and multivariate regression models. Finally, we present the conclusion of this research.

3.2. MATERIAL AND METHODS

3.2.1. DATA SET

The data set used in our analysis were collected at Xanxerê municipality, Santa Catarina, Brazil. The study region belongs to Atlantic Forest, under domain of Mixed Ombrophilous Forest (MOF). The native forest has about 400 hectares of total area. The forest fragment is currently being enriched with *Ilex paraguariensis* A. St.-Hill specie for the purpose of yielding Erva-mate.

The trees were randomly selected on the forest fragment. The data set were composed by 169 independent sample trees. The individuals presented a large variation on their dimensions. The variables diameter at breast height (D, in centimeters – cm), total height (H, in meters – m) and individual volume with bark (V, in cubic meters – m^3) were measured in each sample tree. The tree volume in each section was calculated by Huber's method. The total volume with bark was obtained by summing the partial volumes with the tree top volume (MACHADO & FIGUEIREDO FILHO, 2009).

The tree circumference was measured with measuring tape at the basal portion of the tree, and later converted to a diameter. These sections had shorter length with range of 0.1 – 0.3 m; 0.3 – 0.5 m; 0.5 – 0.7 m; 0.7 – 0.9 m; and 0.9 – 1.3 m. After, the sections had length of 2 m, where the Bitterlich's Spiegel-Relaskop (narrow band) was used for measuring in an indirectly way the diameters on the upper portion of the tree stem. Thus, the measures were taken at the height of 0.2 m; 0.4 m; 0.6 m; 0.8 m; 1.1 m; 2.3 m; and each 2 m until the tree total height.

3.2.2. MULTIVARIATE COVARIANCE GENERALIZED LINEAR MODEL

The modeling process of forest variables by multivariate regression models is not a common approach in forestry research. Therefore, this subsection aimed to present the multivariate covariance generalized linear model structure (MCGLM). This class of models require a specification of the expected value and variance of the response variables (BONAT & JØRGENSEN, 2016). Thus, a generic formulation for the multivariate case is given as

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mathcal{M}} = \{\mathrm{g}_1^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1), \ldots, \mathrm{g}_\mathrm{R}^{-1}(\mathbf{X}_\mathrm{R}\boldsymbol{\beta}_\mathrm{R})\}$$

$$\mathrm{Var}(\mathbf{Y}) = \mathbb{C} = \boldsymbol{\Sigma}_\mathrm{R} \otimes \boldsymbol{\Sigma}_\mathrm{b},$$

where: $\mathbf{Y}_{\mathrm{N}\times\mathrm{R}} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_\mathrm{R}\}$ is a matrix of response, being N the number of observations and R the number of response variables; $\boldsymbol{\mathcal{M}}_{\mathrm{N}\times\mathrm{R}} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_\mathrm{R}\}$ is a matrix of expected values; $\mathbf{X}_\mathrm{R}$ is an N × K design matrix, being *K* the number of covariates; $\boldsymbol{\beta}_\mathrm{R}$ is a $\mathrm{K}_\mathrm{R} \times 1$ matrix of regression vectors; $\mathrm{g}_\mathrm{R}(\cdot)$ is a twice differentiable and monotonous link function; $\boldsymbol{\Sigma}_\mathrm{R} \otimes \boldsymbol{\Sigma}_\mathrm{b} = \mathrm{Bdiag}(\widetilde{\boldsymbol{\Sigma}}_1, \ldots, \widetilde{\boldsymbol{\Sigma}}_\mathrm{R})(\boldsymbol{\Sigma}_\mathrm{b} \otimes \mathbf{I})\mathrm{Bdiag}(\widetilde{\boldsymbol{\Sigma}}_\mathbf{1}^\mathbf{T}, \ldots, \widetilde{\boldsymbol{\Sigma}}_\mathbf{R}^\mathbf{T})$ is the generalized Kronecker product; $\boldsymbol{\Sigma}_\mathrm{R}$ is an N × N covariance matrix within response $\mathrm{r} = 1, \ldots \mathrm{R}$; $\boldsymbol{\Sigma}_\mathrm{b}$ is an *R* × R correlation matrix among response variables; $\widetilde{\boldsymbol{\Sigma}}_\mathrm{R}$ is a lower triangular matrix of Cholesky decomposition of $\boldsymbol{\Sigma}_\mathrm{R}$; operator

Bdiag represent a diagonal block matrix; and **I** is an R × R identity matrix.

The covariance matrix $\boldsymbol{\Sigma_r}$ for each response is given as

$$\boldsymbol{\Sigma}_r = V(\boldsymbol{\mu_r}; \mathbf{p}_r)^{1/2} h\{\boldsymbol{\Omega}(\boldsymbol{\tau}_r)\} V(\boldsymbol{\mu_r}; \mathbf{p}_r)^{1/2},$$

where: $V(\boldsymbol{\mu_r}; \mathbf{p}_r) = \text{diag}\{\vartheta(\boldsymbol{\mu_r}; \mathbf{p}_r)\}$ is a diagonal matrix, whose main entries denote the variance functions $\vartheta(\boldsymbol{\mu_r}; \mathbf{p}_r)$ applied elementwise to the response vector $\boldsymbol{\mu_r}$; $\mathbf{p}_r$ is a power parameter vector; $\boldsymbol{\tau}_r$ is a dispersion parameter vector; $h\{\boldsymbol{\Omega}(\boldsymbol{\tau}_r)\} = \tau_0 \mathbf{Z_0} + \ldots + \tau_T \mathbf{Z_T}$; h is a covariance link function; Z_d are known matrices that describe the covariance structure with $t = 0, \ldots, K$; and τ_T is a $(T + 1) \times 1$ parameter vector. The structure that specify the mean, $E(\mathbf{Y})$, is called linear or non-linear predictor, while the structure that specify the covariance, $\text{Var}(\mathbf{Y})$, is known as matrix linear predictor.

The g link function connect the linear prediction with the expected values of response variable. Appropriated choices of link functions allow to ensure suitable values for the mean. In a similar way, the covariance link function h connect the matrix linear predictor with the covariance of response variable.

The variance function is a fundamental component of the MCGLM. Different assumptions about the distribution of response variable can be performed for different values of variance function (BONAT & JØRGENSEN, 2016). The power parameter of the variance function $\vartheta(\cdot; \mathbf{p_r}) = \boldsymbol{\mu}_r^{\mathbf{p_r}}$ characterize the Tweedie distribution family, and the most important special cases are Normal ($p = 0$), Poisson ($p = 1$), composed Poisson-Gamma ($1 < p < 2$), Gamma ($p = 2$) e Inverse Normal ($p = 3$) distributions.

3.2.3. STATISTICAL ANALYSIS OF THE DATA

MCGLM were used for modeling the variables measured on the *Araucaria angustifolia*. The response variables were tree height (H) and volume (V), which both are continuous. The diameter was used as the only covariate, also continuous, and its effect varied until third degree.

At the beginning, we fit just univariate regression models. This means that both responses H and V were considered as independent and the models were treated in a separately way. Then, our models were jointly fitted in a multivariate approach. In this context, the focus was to analyze the influence of correlation between response variables on the point estimates

and standard errors of the fitted models.

Both response variables have an apparently non-constant variance over time. This feature is related to the natural process of tree growth. Thus, it is expected that the variance of the observations increases according to the dimensions of the individuals. For modeling the non-constant variance pattern, we tested two approaches: I) the variance modeling was performed directly on the matrix linear predictor as a linear function of covariate D, and the effects were represented by a third-degree polynomial; II) a variance function $\vartheta(\boldsymbol{\mu}_r; \mathbf{p}_r)$ was specified on the matrix linear predictor, which allowed a directly modeling of mean and variance of the response variable.

The performance of the models was compared by a gaussian pseudo likelihood (PL), and a pseudo Bayesian's information criterion (PBIC). The PV is a similar measure to the log-likelihood value from the maximum likelihood estimation context. Therefore, the highest value of PV suggests the best fitted model. The PBIC has an advantage to penalize the models with higher number of parameters, and lowest value suggest the best fit (BONAT, 2018).

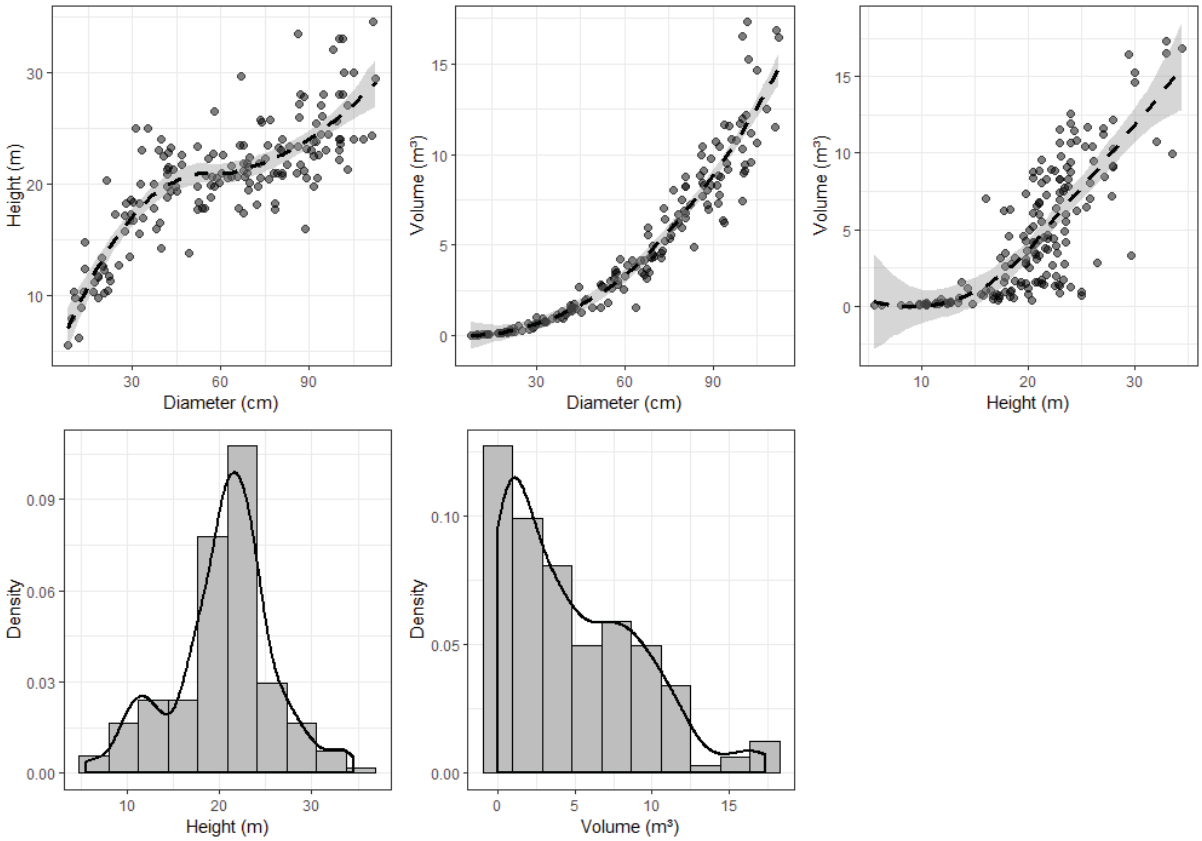
The fit of the univariate and multivariate regression models was performed on the R statistical software (R CORE TEAM, 2019) by using the mcglm package, version 0.5.0 (BONAT, 2018). The package has an intuitive interface and many functions are available for building the components of the matrix linear predictor and fitting the regression models. The ggplot2 package was also used for building the graphics (WICKHAM, 2016).

3.3. RESULTS AND DISCUSSION

3.3.1. EXPLORATORY DATA ANALYSIS

Exploratory data analysis was performed in order to understand the behavior of the response variables height and volume, and their relationship with the covariate diameter. Histograms presented in FIGURE 3.1 suggested that the response variables did not have the same probability distribution. Variable volume V presented a strongly asymmetric distribution, what suggested to include a logarithmic link function on the linear predictor of the regression models. Still, we can note that the relationship between tree average volume and variance was not constant for different diameters, and a tendency to increase the variability for larger trees was empirically observed.

FIGURE 3.1 – HISTOGRAMS OF RESPONSE VARIABLES HEIGHT (H) AND VOLUME (V) AND SCATTER PLOT BETWEEN RESPONSE VARIABLES AND COVARIATE DIAMETER (D). SOLID LINE IN BLACK COLOR IS A KERNEL DENSITY ESTIMATE. DASHED LINE IN BLACK COLOR IS A LOCALLY ESTIMATED SCATTERPLOT SMOOTHING WITH 95% CONFIDENCE INTERVALS



These initial results suggested the inclusion of a Tweedie variance function and a logarithmic link function, once that characterize many continuous asymmetric distributions, and allows to model mean and variance relationship of response variable V in a suitable way. By the other side, the response variable H presented an apparently symmetric distribution, indicating an identity link function on the linear predictor, besides a constant variance function, what is an assumed assumption for variables with normal distribution.

In this research, our main interest was to model the components of the matrix linear predictor. Therefore, linear prediction was specified in a preliminary data analysis, and the same structure was used for univariate and multivariate fitting. The expected value of observation i of response H_i was specified from the cubic effect of covariate D_i, while the expected value of V_i was specified based on quadratic effect of D_i. Thus, the linear predictor with a link function for both responses were given as

$$E(H_i) = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 D_i^3,$$

$$E(V_i) = \exp(\beta_0 + \beta_1 D_i + \beta_2 D_i^2).$$

In addition to the variance function used for modeling the mean and variance relationship of response variable, which was estimated from a power parameter, four structure were tested for variance modeling on the matrix linear predictor. Thus, the linear combination of known matrices applied for both response was given as

$$M1 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I,$$

$$M2 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z,$$

$$M3 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z + \tau_2 Z^2,$$

$$M4 = \boldsymbol{\Omega}(\boldsymbol{\tau}_r) = \tau_0 I + \tau_1 Z + \tau_2 Z^2 + \tau_3 Z^3,$$

where: I is an N × N identity matrix, being N the number of observations; Z is an N × N diagonal matrix whose main entries are constituted by tree diameters (D), where its effects varied until third degree.

The main results from univariate and multivariate fitting for response variables height and volume are given on the next topics.

3.3.2. UNIVARIATE MODELS

Parameter estimates and standard errors of the univariate models for response H and V are presented in TABLES 3.1 e 3.2, respectively. In general, the intercepts of the linear predictors were not significant for H at 5% level (fixed in all analyzes), while the other estimated parameters were significant. Dispersion parameters were significant, but negative effects were observed for $\hat{\tau}_1$ of the M3 model and $\hat{\tau}_2$ of the M4 model for the response variable H.

Parameter estimates of the linear predictor were significant for response V, while dispersion parameters $\hat{\tau}_0$ and $\hat{\tau}_1$ did not for M2 model. When we considered the response V,

estimated power parameter did not differ from 1.5 for the M1 model. This value characterizes a composed Poisson-Gama distribution from Tweedie family, indicating the variance of the data increased faster than mean values.

The non-significance of the power parameter of the M2 model indicated a Normal distribution ($p = 0$), what suggested that the variance is always constant. Finally, the confidence intervals of power parameter from the M3 model suggested a Poisson distribution ($p = 1$), indicating the variance increase proportional to mean values. Still, the M4 model did not present a convergency, even when the estimation algorithm was relaxed in order to facilitate the fitting.

TABLE 3.1 – POINT ESTIMATES AND STANDARD ERRORS OF THE UNIVARIATE MODELS FOR RESPONSE VARIABLE HEIGHT (H)

Model	Parameter	Estimates	Standard error
M1	β_0	0.3776	1.7087
	β_1	0.8937	0.1114
	β_2	-0.0131	0.0021
	β_3	0.00007	0.00001
	τ_0	8.7814	1.1021
M2	β_0	0.1853	1.4851
	β_1	0.9082	0.1019
	β_2	-0.0133	0.0020
	β_3	0.00007	0.00001
	τ_0	5.0865	1.8687
	τ_1	0.0600	0.0340
M3	β_0	0.0205	1.6592
	β_1	0.9235	0.1095
	β_2	-1.3679	0.002077
	β_3	0.00007	0.00001
	τ_0	9.3133	2.6558
	τ_1	-0.1236	0.1174
	τ_2	0.0015	0.0011
M4	β_0	0.3125	1.3065
	β_1	0.9053	0.0962
	β_2	-1.1340	0.0019
	β_3	0.00007	0.00001
	τ_0	-2.4738	2.3795
	τ_1	0.7560	0.2815
	τ_2	-0.0164	0.0065
	τ_3	0.0001	0.00004

TABLE 3.2 – POINT ESTIMATES AND STANDARD ERRORS OF THE UNIVARIATE MODELS FOR RESPONSE VARIABLE VOLUME (V)

Model	Parameter	Estimates	Standard error
M1	β_0	-3.1707	0.1108
	β_1	0.0988	0.0036
	β_2	-0.0004	0.00003
	τ_0	0.0830	0.0102
	p	1.6350	0.0829
M2	β_0	-3.1614	0.1110
	β_1	0.0986	0.0036
	β_2	-0.0004	0.00003
	τ_0	0.0465	0.3032
	τ_1	0.0009	0.0083
	p	1.4389	1.5765
M3	β_0	-2.9687	0.1131
	β_1	0.0939	0.0037
	β_2	-0.0004	0.00003
	τ_0	0.0742	0.0264
	τ_1	-0.0036	0.0017
	τ_2	0.00008	0.00005
	p	0.9376	0.3377

3.3.3. MULTIVARIATE MODELS

Parameter estimates and standard errors of the univariate models for responses H and V are presented in TABLES 3.3 e 3.4, respectively. Just M1 model for response H presented significance in all parameters at 5% confidence level. However, M1 and M3 model for response V presented all parameters significant; and the power parameters were similar to that ones from univariate fitting. We also observed negative effects for dispersion parameter $\hat{\tau}_1$ of the M3 model for response V. The M4 model did not converge in a similar way that the univariate case.

As we previously mentioned, MCGLM are based on a marginal specification, i.e., the models require a specification of the mean and variance components. The marginal specification of our models has the advantage of not testing the null hypothesis of the dispersion parameters $\tau = 0$ on the border of parametric space, such as the likelihood ratio, Wald and score tests (BONAT, 2017). Thus, it is common to observe negative effects from dispersion parameters for the variance components, like that ones we observed in this research.

TABLE 3.3 – POINT ESTIMATES AND STANDARD ERRORS OF THE MULTIVARIATE MODELS FOR RESPONSE VARIABLE VOLUME (H)

Model	Parameter	Estimates	Standard error
M1	β_0	5.7281	1.5568
	β_1	0.5201	0.0994
	β_2	-0.0061	0.0018
	β_3	0.00003	0.00001
	τ_0	9.3747	1.1124
M2	β_0	5.1783	1.4466
	β_1	0.5543	0.09526
	β_2	-0.0067	0.0018
	β_3	0.00003	0.00001
	τ_0	7.0741	2.0709
	τ_1	0.0353	0.0366
M3	β_0	4.4628	1.4632
	β_1	0.6141	0.0968
	β_2	-0.0079	0.0018
	β_3	0.00004	0.00001
	τ_0	7.6747	1.6314
	τ_1	-0.0036	0.0867
	τ_2	0.0003	0.0010

TABLE 3.4 – POINT ESTIMATES AND STANDARD ERRORS OF THE MULTIVARIATE MODELS FOR RESPONSE VARIABLE VOLUME (V)

Model	Parameter	Estimates	Standard error
M1	β_0	-3.1528	0.1122
	β_1	0.0982	0.0036
	β_2	-0.0004	0.00003
	τ_0	0.0843	0.0095
	p	1.6152	0.0781
M2	β_0	-3.1541	01116
	β_1	0.0984	0.0036
	β_2	-0.0004	0.00003
	τ_0	0.0426	0.2130
	τ_1	0.0010	0.0059
	p	1.4080	1.0991
M3	β_0	-3.0332	0.1113
	β_1	0.0953	0.0036
	β_2	-0.0004	0.00003
	τ_0	0.0527	0.0063
	τ_1	-0.0034	0.0014
	τ_2	0.0001	0.00004
	p	0.7738	0.2631

Univariate and multivariate models present the same interpretations about the estimated effects of the linear predictor and matrix linear predictor. Differences observed on the parameter estimates and standard errors from univariate and multivariate approaches were due to the correlation between both response variables, as showed in TABLE 3.5. Correlations ($\hat{\rho}$) between response variables H and V was significant for all models, but in a moderate intensity with values close to 0.5, indicating that these variables share information. As consequence, standard errors from multivariate model was smaller than univariate models for almost all estimated parameters.

TABLE 3.5 – ESTIMATED CORRELATION ($\hat{\rho}$) BETWEEN RESPONSE VARIABLES HEIGHT (H) AND VOLUME (V) FOR THE MULTIVARIATE FITTING

Model	Estimate	Standard error
M1	0.5075	0.0595
M2	0.4945	0.0604
M3	0.4715	0.0619

3.3.4. PERFORMANCE OF THE FITTED MODELS

The performance of the univariate and multivariate models for response variables H and V are presented at Table 3.6. To make the performance measure comparable, we sum the values from univariate fitting for the same response. Pseudo likelihood (PV) values tended to increase due to the inclusion of news parameters on the matrix linear predictor, suggesting best performance for the most parametrized models. However, pseudo Bayesian's information criterion (PBIC) penalized models as a function of the number of parameters.

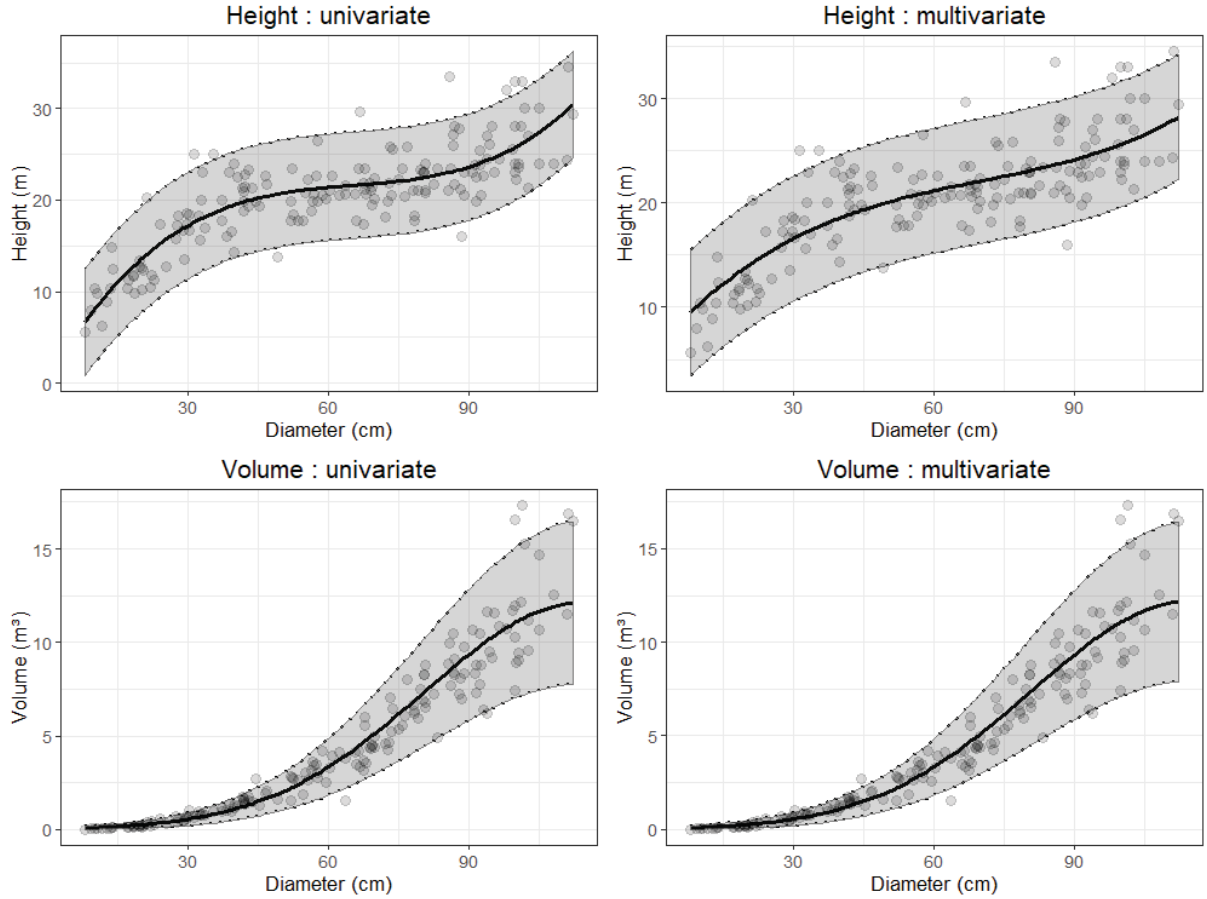
The best fitted models were obtained from the simplest models, where the matrix linear predictor was composed of only an identity matrix, with an estimated dispersion parameter common for all observations. However, we highlighted that the variance function estimated from jointly modeling for response variable V suggested a composed Poisson-Gama distribution. These results indicated that it is important to include components for variance modeling for the response V, since it was not constant and tended to increase faster than the mean values. Still, we can note that the variance function was able to handle with the natural variability of the data, avoiding the need to model the variance as function of covariate diameter (D) and to specify a more complex model, as we performed on models 2 to 4.

The M1 specification was the most suitable for modeling the behavior of response variables for both univariate and multivariate approaches, mainly due to its simplified formulation combined to the lower values of PBIC. From this model specification, 95% confidence intervals were built for response variables at FIGURE 3.2. We can note that the confidence intervals were symmetric for response H, while larger confidence intervals were observed for response V for larger trees, due to the variance function, being slightly lower for multivariate fitting.

TABLE 3.6 – PSEUDO LIKELIHOOD (PV) AND PSEUDO BAYESIAN'S INFORMATION CRITERION (PBIC) FROM UNIVARIATE AND MULTIVARIATE MODELS FOR RESPONSE VARIABLE HEIGHT (H) AND VOLUME (V)

Model	PV	PBIC
Model for H and V: univariate case		
M1	-570.57	1199.31
M2	-567.80	1205.41
M3	-564.70	1210.84
Model for H and V: multivariate case		
M1	-550.98	1165.95
M2	-549.95	1175.52
M3	-548.09	1183.44

FIGURE 3.2 – 95% CONFIDENCE INTERVALS FROM UNIVARIATE AND MULTIVARIATE MODELS FOR RESPONSE VARIABLES HEIGHT (H) AND VOLUME (V)



3.4. CONCLUSION

Univariate and multivariate regression models fitted for describing the response variables height and volume of *Araucaria angustifolia* specie were suitable. The correlation between response variables can influence the parameter estimates and standard errors of the fitted models. The variance function has potential to improve the performance of the models for both approaches and allows a suitable variance modeling of the response variables. Thus, multivariate covariance generalized linear models are a class of models with great potential to be applied to forest biometric for estimating tree-level variables.

REFERENCES

BONAT, H.W. Modelling Mixed Types of Outcomes in Additive Genetic Models. **The international journal of biostatistics**, v. 13, n. 2, p. 1–16, 2017.

BONAT, W.H. Multiple response variables regression models in R: The mcglm package. **Journal of Statistical Software**, v.84, n.4, p.1-30, 2018.

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society. Series C: Applied Statistics**, v. 65, n. 5, p. 649–675, 2016.

BONAT, W.H.; OLIVERO, J.; GRANDE-VEJA, M.; FARFÁN A.; FA, J.E. Modelling the Covariance Structure in Marginal Multivariate Count Models: Hunting in Bioko Island. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 22, n. 4, p. 446–464, 2017.

FU, L.; ZHANG, H,; SHARMA, R.P.; PANG, L.; WANG, G. A generalized nonlinear mixed-effects height to crown base model for Mongolian oak in northeast China. **Forest Ecology and Management**, v. 384, p. 34–43, 2017.

LAM, T.Y.; KERSHAW, J.A.; HAJAR, Z.S.N.; RAHMAN, K.A.; WEISKITTEL, A.R.; POTTS, M.D. Evaluating and modelling genus and species variation in height-to-diameter relationships for Tropical Hill Forests in Peninsular Malaysia. **Forestry**, v. 90, n. 2, p. 268–278, 2017.

LAPPI, J.A. multivariate, nonparametric stem-curve prediction method. **Canadian Journal of Forest Research**, v. 36, n. 4, p. 1017–1027, 2006.

MACHADO, S.A.; FIGUEIREDO FILHO, A. **Dendrometria**. Guarapuava: UNICENTRO, 2009.

MARCHIORI, J.N.C. **Dendrologia das gimnospermas.** Editora ufsm, ed. 2, 2005.

MEHTÄTALO, L.; DE-MIGUEL, S.; GREGOIRE, T.G. Modeling height-diameter curves for prediction. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 826–837, 2015.

R CORE TEAM. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, 2019.

SCHEEREN, L.W.; FINGER, C.A.G.; SCHUMACHER, M.V.; LONGHI, S.L. Crescimento em altura de *Araucaria angustifolia* (Bert.) O. Ktze. em três sítios naturais, na região de Canela (RS). **Ciência Florestal**, v. 9, n. 2, p. 23–40, 1999.

THOMAS C.; ANDRADE, C.M.; SCHNEIDER, P.R.; FINGER, C.A.G. Comparação de equações volumétricas ajustadas com dados de cubagem e análise de tronco. **Ciência Florestal**, v. 16, n. 3, p. 319–327, 2006.

WEBER, K.S.; SANQUETA, C.R.; EISFELD, R.L. Variação volumétrica e distribuição espacial do estoque de carbono Em Floresta Ombrófila Mista. **Revista Acadêmica: Ciência**

Animal, v. 3, n. 2, p. 77–85, 2017.

WICKHAM, H. **Ggplot 2: elegant graphics for data analysis**. Springer, 2016.

4. MODELOS LINEARES GENERALIZADOS PARA SOBREVIVÊNCIA DE *Pinus taeda* EM PLANTIOS FLORESTAIS

RESUMO

Quantificar a sobrevivência de árvores em povoamentos florestais e estimar a probabilidade uma árvore sobreviver são questões fundamentais no planejamento do manejo florestal. Portanto, o principal objetivo da pesquisa é estimar a probabilidade de sobrevivência de árvores em plantios de *Pinus taeda* baseado nos modelos lineares generalizados (GLM). O conjunto de dados foi obtido em inventários florestais realizados na região Meio-Oeste do estado de Santa Catarina, Brasil. A análise dos dados combinou estratégias para seleção de covariáveis e diferentes especificações de funções de ligação no modelo GLM Bernoulli. As estratégias para seleção de covariáveis a nível de parcela foram o procedimento Stepwise tradicional, além da abordagem elastic net, bem como os seus casos particulares de penalização lasso e ridge. A análise mostrou que o procedimento stepwise, combinado com a função de ligação complemento log-log proporcionou o melhor ajuste. As variáveis que mais contribuíram para predizer a sobrevivência das árvores foram área basal, número de indivíduos, diâmetro máximo, diâmetro médio da área transversal média e o coeficiente de variação dos diâmetros por parcela. Esse modelo apresentou 81,5% de acurácia, conforme a curva ROC. Por fim, o modelo ajustado também foi avaliado pelo gráfico half-Normal plots e os resíduos quantílicos aleatorizados, os quais indicam um ajuste adequado do modelo. O procedimento Stepwise é recomendado para selecionar covariáveis para predizer a probabilidade de sobrevivência de árvores, juntamente com uma função de ligação complemento log-log.

Palavras-chave: Elastic net. Função de ligação, Regressão logística. Regressão ridge. Método stepwise.

GENERALIZED LINEAR MODELS FOR TREE SURVIVAL IN LOBLOLLY PINE PLANTATIONS

ABSTRACT

To quantify the surviving trees in a forest stand and estimate the probability of an individual tree to survival are a fundamental task in forest management planning. Therefore, the main goal of this paper was to estimate the tree survival probability in loblolly pine (*Pinus taeda* L.) plantations based on generalized linear models (GLM). The data set was obtained from forest inventories carried out in the Midwest of Santa Catarina State, Brazil. The data analysis combined strategies for selecting covariates and different specifications of link functions in a Bernoulli GLM. We performed strategies for covariate selection at plot-level along with the standard stepwise procedure, where we considered the elastic net approach, as well as its special cases the lasso and ridge penalization. Our analyses showed that the stepwise procedure combined with the complementary log-log link function provide the best fit. The variables that most contributed to assess tree survival were basal area, number of individuals, maximum diameter, diameter of the average cross-sectional area and the diameter coefficient of variation per plots. This model presents 81.5% of accuracy given by ROC curve. Finally, we evaluated the fitted model by means of the half-Normal plots and randomized quantile residuals, whose results showed evidence of a suitable fit. We suggest the stepwise procedure for selecting covariates for a tree survival probability model, besides a complementary log-log link function.

Keywords: Elastic net. Link function. Logistic regression. Ridge regression. Stepwise method.

4.1. INTRODUCTION

Species of Pinus genus are cultivated in large-scale in the Southern region of Brazil, especially in Paraná and Santa Catarina States, mainly due to great adaptation to climatic conditions and their high timber economic potential. According to IBÁ - Indústria Brasileira de Árvores (2017), *Pinus taeda* L. and *Pinus elliottii* Engelm planted area covered more than 1.6 million of hectares in the base year of 2016, which represent 20.4% of the total planted area in the country.

The extensive Pinus planted area in Brazil implies that the trees are submitted to a wide range of environmental conditions and forest management systems, which results in a large range of timber productions. Therefore, statistical models able to express the forest developing in different conditions has become an important tool on the growth and yield planning. Despite important in individual tree growth simulators, tree survival is still few explored, probably because it is a rare phenomenon of high variability (AVILA & BURKHART, 1992).

The tree survival and mortality in both planted and natural forest stands is a phenomenon associated to many factors (ADAME et al., 2010) which include the competition among individuals; forest management practices, such as thinnings; climatic conditions (DIÉGUEZ-ARANDA et al., 2005; DAS & STEPHENSON, 2015; THAPA & BURKHART, 2015, MIRANDA et al., 2017; TÉO, 2017); as well as the species genetic diversity. Thus, it is not completely clear how the tree mortality or survival occurs in a forest, once that individuals with similar features may present different outcomes.

To quantify the number of surviving trees over time is important in forest plantations. This information indicates the number of trees expected in the silvicultural rotation; and the potential timber assortments for being explored on the industry. Based on this, the tree survival probability at different site conditions and management systems can be obtained by statistical tools. Furthermore, regression models are essential on forest planning because can assist to identify factor associated to high or low survival probabilities.

One of these tools is logistic regression, a statistical approach widely used for estimating the tree survival probability in forest plantations (YAO et al., 2001; DIÉGUEZ-ARANDA et al., 2005; THAPA & BURKHART, 2015; TÉO, 2017). This model allows to express the survival probability through a linear predictor, which is usually composed by a set of tree and plot-level covariates. The linear predictor is connected to the expectation of the

survival probability by a link function, frequently specified by a logit or probit functions (TÉO, 2017; VANCLAY, 1991; YANG et al., 2003; YAO et al., 2001).

Although their popularity, both logit, probit and Cauchit link functions share a limitation for the reason they are symmetric (McCULLAGH & NELDER, 1989). In practice, this feature can be a limitation depending of the data sets. Thus, complementary log-log asymmetric link functions are available in the statistical literature as alternative approaches. However, the suitability of these link functions is not well-known in the context of forest management research, doing this subject quite relevant.

The specification of a statistical model for modeling tree survival has at least two crucial choices: a suitable link function and which covariates will compose the linear predictor. In general, forest researchers have been used forward, backward, or stepwise selection procedures, where satisfactory results have been reported (TÉO, 2017; ZHANG et al., 2017). In this paper, we introduce an alternative approach for selecting covariates at plot-level based on regularization methods. The main idea of these methods is to fit a regression model whose parameter estimates are penalized or shrunken toward to zero. In this approach, the goal is to obtain estimates with lower variance at the cost of introducing some bias in the parameter estimates. This feature of the regularization methods can be used for selecting covariates measured in forest plantations, once that they present high value of correlation among them, which implies in large standard errors.

Lasso (Least Absolute Shrinkage and Selection Operator) and Ridge regression are a frequently applied regularization technique (TIBISHIRANI, 1996). An extension of these strategies is the Elastic Net approach which is a combination of Lasso and Ridge penalizations (ZOU & HASTIE, 2005). These techniques penalize the covariates by shrinking the parameter estimates and enabling the removal of the covariates whose estimated effects approach zero. Thus, our research hypothesis is that the regularization methods are appropriated for selecting uncorrelated covariates or non-redundant, once this approach can reduce the variance of the parameter estimates.

The aim of this paper was to estimate the probability of loblolly pine (*Pinus taeda*) survival in forest plantations; and to identify which factors are associated to the tree survival. Therefore, we obtained data from forest inventories carried out in the Midwest of Santa Catarina State, Brazil. Our data was composed by a set of covariates usually measured at plot-level. The response variable was a binary value that indicated whether the tree is alive or not. We investigated and compared Bernoulli's regression models fitted by the link functions logit, probit, Cauchit and complementary log-log. Furthermore, we performed the covariates

selection based on the standard stepwise procedure, as well as the methods based on regularization as the Lasso, ridge and elastic net approaches.

We present the data set, a brief description about the study area and the modeling strategies in the material and methods section. The results section describes an exploratory data analysis and the application of the models to the data. We also present a discussion section about the main results. Finally, concluding remarks are presented in the conclusion section.

4.2. MATERIAL AND METHODS

4.2.1. STUDY AREA

The study area corresponds to loblolly pine (*Pinus taeda*) plantations, located at Midwest region of Santa Catarina state, Brazil. The plantations are distributed on the municipalities of Caçador, Lebon Régis, Macieira, Rio das Antas, Santa Cecília, and Timbó Grande. According to IBGE - Instituto Brasileiro de Geografia e Estatística (2012), the region presents original vegetation belonging to Mixed Ombrophilous Forest (MOF), under the Montane Mixed Ombrophilous Forest. Based on the Köppen classification, the study region presents a Cfb climate type, that is, a wet subtropical zone, oceanic climate, without a dry season and with summers temperate. The average temperature of the warmest month is 19.7 °C and the coldest month is 11.5 °C, and the annual precipitation is 1,736 mm (ALVARES et al., 2013).

The forest plantations were planted with an average initial spacing of 2.5 x 2.5 m (1,600 trees per hectare). The ideal rotation age is 25 years, with three commercial thinnings usually performed at 10, 15, and 20 years old. In the first thinning, 50% of the trees per hectare were removed; while 40% of the remaining trees were removed in the second thinning; in the third and last thinning were removed 30% of the remaining trees.

4.2.2. DATA SET

The data set was obtained from forest inventory performed in two occasions carried out at 2009 and 2015. The age ranged from 5.5 to 35.2 years old. In addition, due to the difference of six years between both forest inventories and because we re-sampling a few sample units, we assume that there is no correlation between measures.

The plots had dimensions from 497,93 to 739,68 m^2, which were randomly allocated (simple random sampling) in the study area, by using a stratified sampling process. The stratum

represented administrative divisions of the company (projects and stands). The diameter at breast height (DBH) was measured at 1.30 m of height in all trees inside each sub-sample. The total height of 20% of the trees in each plot was indirectly taken by using a hypsometer Vertex III. The trees dominant height was measured in individuals without bifurcation or defects over the stem and crown, and it was defined proportionally as the 100 trees with largest diameter at breast height per hectare.

The data set we used for modeling was composed by 13 random variables measured at plot-level. The number of trees selected was 40,556 trees. The description of each variable is given as follow:

- *survival*: binary variable – takes value 1 if the tree is alive or 0 otherwise. The classification of alive tree was performed when the data was collected in the forest inventory. The tree was considered a dead individual when green branches were not observed on the field. In our approach, both regular and irregular mortality were combined. The regular mortality was due to the natural competition among trees and the senescence process. The irregular mortality was caused by irregular factors, as monkey attacks, which are quite common at the study area.
- *age*: continuous variable – age of the tree (years);
- *gsample*: continuous variable – sum of cross-section areas (m²) of the trees inside plot.
- *nsample*: discrete variable – number of trees inside plot;
- *daverage*: continuous variable – average diameter (cm) of the trees inside plot;
- *dcv*: continuous variable – coefficient of variation (%) of the diameters inside plot;
- *dg*: continuous variable – quadratic average diameter (cm) of the trees inside plot;
- *dmax*: continuous variable – maximum diameter (cm) of the trees inside plot;
- *ddom*: continuous variable – dominant diameter (cm) of the trees inside plot. This variable was computed based on the average diameter of the one hundred largest trees per hectare, but proportionally to the size of each plot;
- *hdom*: continuous variable – dominant height (m) of the plot. This variable was computed based on the height average of the one hundred largest trees by hectare, but proportionally to the size of each plot;
- *thinsample*: binary variable – takes value 1 whether were performed thinnings on the plot or 0 otherwise;
- *gthin*: continuous variable – sum of removed cross-sectional area on the plot during the thinnings;
- *nthin*: discrete variable – number of trees removed during the thinnings on the plot.

4.2.3. THE GENERALIZED LINEAR MODEL

Tree survival (survival) was the response variable, which takes a binary value, i.e., the response variable take value 1 whether the tree is alive and 0 otherwise. Therefore, we applied a Bernoulli's regression model due to the nature of response variable (McCULLAGH & NELDER, 1989). The systematic component was formulated by a linear combination of a set of predictor variables, besides a link function selected according to the behavior of response variable. The specification of the model is given as

$$Y_i|x_i \sim Bernoulli(\pi_i) \text{ and}$$

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

where Y_i is the random variable, whose observed values are denoted by y_i, $i = 1,2,\ldots,n$; $x_{i1},\ldots,x_{ip}$ are vectors of the predictor variables X_i; π_i is the probability of success, i.e., is the survival probability; g is a differentiable and monotone link function; η_i is the linear predictor; and $\beta_0,\beta_1,\ldots,\beta_p$ are parameters to be estimated.

4.2.4. LINEAR PREDICTOR AND LINK FUNCTION SELECTION

For composing the linear predictor, 12 covariates were available. We applied two strategies for selecting the covariates:

I) Stepwise: covariate selection was based on the minimization of the Bayesian Information Criterion (BIC), given by the following expression

$$BIC = -2\hat{l} + ln(n)p,$$

where $\hat{l}$ is the maximized log-likelihood value; n is the number of observations; and p is the number of parameters of the model. This algorithm is a combination of backward and forward procedure, where the covariates are added or removed in successive iterations until obtaining the smallest BIC. Thus, we assumed that this methodology is the standard approach in forest modeling due to its large applications.

II) Regularization: covariates selection was performed with regularization methods. This procedure is based on penalizations controlled by the parameter λ; while the penalization intensity was quantified by parameter α. The general formulation is given by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{l}\left(y_i, \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) + \lambda\left[\alpha \sum_{i=1}^{n} ||\beta_p|| + (1-\alpha)\sum_{i=1}^{n} ||\beta_p^2||\right].$$

For the especial case where $\alpha = 1$, we obtained a first order penalization, also called as lasso regularization method. A second order penalization was defined when $\alpha = 0$, and the method is called as ridge regression. The Elastic Net is an intermediate penalization when $0 < \alpha < 1$, and we tested a large grid (one hundred points) of penalization intensity. The optimum λ was determined by cross-validation, using the cv.glmnet function of the glmnet package (FRIEDMAN et al., 2010) on the R software (R CORE TEAM, 2019). In this approach, our main goal was to identify the smallest loss for a sequence of λ. Still, we tested the loss function based on Mean Squared Error (MSE), Mean Absolute Error (MAE) and Deviance (DEV). Once that the $\lambda \geq 0$, the penalization term has no effect when $\lambda = 0$, and the parameter estimates are equal to the maximum likelihood estimates. However, when $\lambda > 0$ the penalization is strong and the parameter estimates tend to zero (TIBSHIRANI, 1996). The covariates have different nature, what can influence on the selection procedure; thus, we standardized them for minimizing their scale effects.

After defining the best α and λ parameters and the covariates selected on the regularized model, we specified four link functions. The Cauchit (1), complement log-log (2), logit (3) and probit (4) link functions were tested for verifying their influence on the selection of covariates for the stepwise approach. The most suitable link function was based on the smallest value of Bayesian Information Criterion (BIC), once that the models can present different number of covariates. The generalized linear model specification for each link function on linear predictor scale is given by

$$tan[\pi_i(\pi_i - 0.5)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad (1)$$

$$ln[-ln(1-\pi_i)] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad (2)$$

$$ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad (3)$$

$$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \quad (4)$$

where tan is the tangent function; ln is the natural logarithm; and ϕ^{-1} is the inverse of probability density function of the standard normal distribution.

Investigating the violation of the assumptions of non-normal models usually cannot be done by traditional residual analysis. Therefore, for evaluating the assumptions of the fitted models, we performed a diagnostic analysis based on half-Normal plots (HNP) with simulated envelope, built by hnp function of the hnp package (MORAL et al., 2017) on the R software. The idea behind HNP is to verify whether the error distribution was specified in an appropriately way. Thus, for a well-fitted model, the simulated envelope is such that the model diagnostics measures are likely to fall within it. The main purpose of the envelope is to serve as an indicative of what we expect about the residuals under a well-fitted model (MORAL et al., 2017). Still, we computed the randomized quantile residuals (RQR) as a complementary analysis. In this case, whether our model is correctly specified, we expect the residuals follow a normal distribution (DUNN & SMYTH, 1996).

4.2.5. PREDICTIVE PERFORMANCE

The predictive performance of the models was compared by standard methods. The data set was randomly split in two subsets. The fitting data was composed by approximately 90% of the observations and was used to fit the models, where the marginal proportion of alive trees was about 97.82%. The validation data set presented 97.80% of alive trees and was applied for evaluating the prediction performance of them models by Receiver Operating Characteristic curve (ROC) of the ROCR package (SING et al., 2005) on the R software. The sensibility (*Sens*) and specificity (*Esp*) of each model was estimated for 0.75; 0.85; 0.90 and 0.95 probability cut points. These measures indicate the performance of the models for classifying individuals in survival or non-survival, in which the more suitable cut point was obtained based on Youden ($Youden$) and Closest Topleft rules (CT, UNAL, 2017), whose expression are given respectively as

$$Youden = \max[Sens + Esp]$$

$$CT = \min\left[(1 - Sens)^2 + (1 - Esp)^2\right].$$

The sensitivity was expressed by the proportion of trees predicted as survivor given the total of alive trees, and allows to quantify the ability of the models in identifying correctly the survival trees; the ability in identifying correctly the dead trees is obtained by the specificity, which it was calculated by the proportion of trees predicted as non-survivor given the total of dead trees. Thus, sensitivity and specificity are measures conditionate to the alive and dead trees observed on the sample, respectively, and both directly depends of the probability cut point. We also computed the positive predictive value (PPV) and negative predictive value (NPV) measures because also depends of the incidence of survival on the forest stands, and both measures are conditionate to the predicted alive and dead trees, respectively (GIOLO, 2017).

4.3. RESULTS

In this section, we presented an exploratory analysis of the variables and how they are related which others. We also showed the effect of the link functions in selecting covariates for composing the linear predictor of the generalized linear model, besides the main results obtained on the covariates selection procedure. Finally, we applied the best models in the validation data set for assessing their prediction performance.

4.3.1. EXPLORATORY DATA ANALYSIS

Boxplots presented in FIGURE 4.1 suggested an asymmetric distribution of the covariates according to the response variable levels and a possible significant effect of the covariates based on *diameters* measures and *age*. FIGURE 4.2 presents a correlogram based on Spearman´s rank correlation coefficient, where the covariates were clustered by the centroid method (MINGOTI, 2005). Three groups with high correlation values stand out, which suggest that multicollinearity can be a concerning problem for this data set and highlights the need of a covariate selection. The *nsample* covariate had a negative relationship with other covariates that directly express tree dimensions. This indicates that as the number of trees in the sample increase, the tree individual dimensions tend to decrease. Moreover, thinnings covariates showed high positive correlation among them, but negative correlation with almost all the other covariates. High positive correlation values were also observed for covariates directly computed based on tree-level measures, such as diameter and height.

FIGURE 4.1 - BOXPLOTS (A TO K) AND BARCHART (L) OF THE COVARIATES BY SURVIVAL INDEX

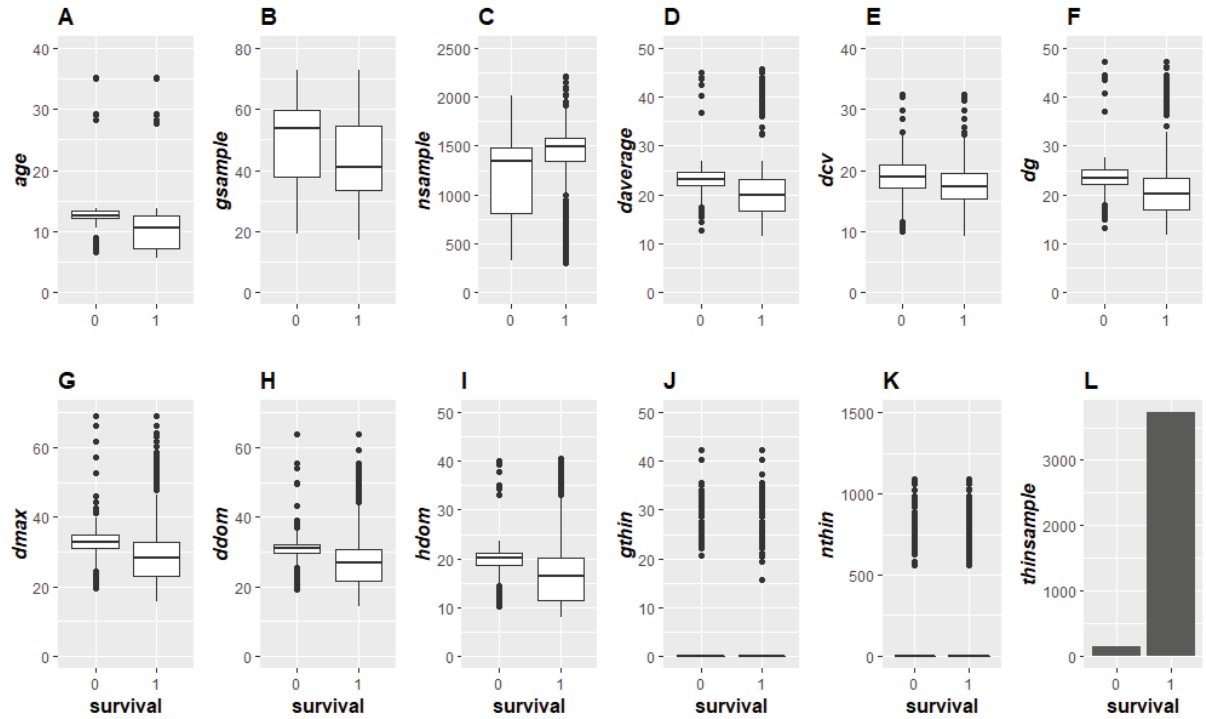
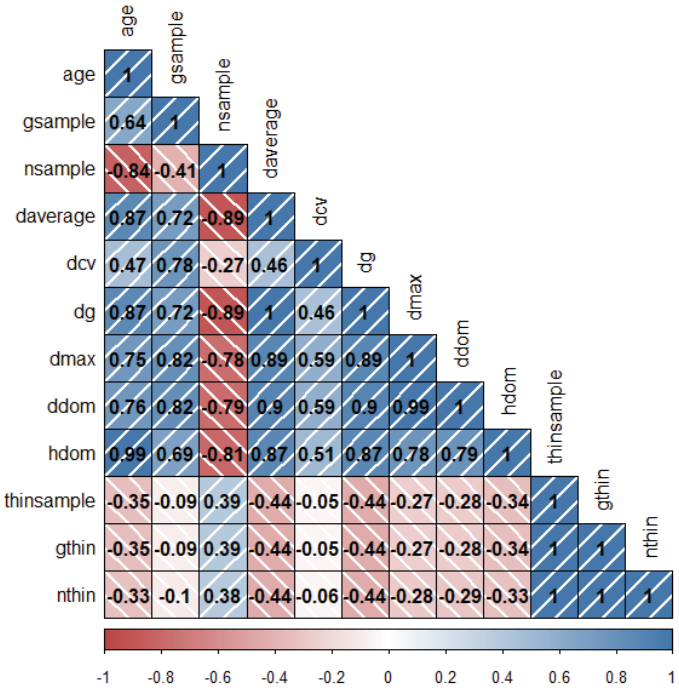


FIGURE 4.2 - CORRELOGRAM BETWEEN VARIABLES CLUSTERED BY CENTROID METHOD



4.3.2. FITTING THE MODELS

The stepwise procedure selected the covariates *gsample*, *nsample*, *dcv*, *dg* and *dmax* for composing the linear predictor. On the other hand, all covariates were selected by the regularization methods. The best λ value obtained by cross-validation was 0 for all sequences of $0 \leq \alpha \leq 1$ that perform the Lasso, Ridge, and Elastic Net procedures, regardless of loss measure tested (MSE, MAE or DEV), indicating that the penalization term had no effect on the parameter estimates, being recommended to remove them. However, even when we fitted the model with all covariates, only *gsample*, *nsample*, *dcv*, *dmax*, *gthin,* and *nthin* were significant (TABLE 4.2). Thus, we decided to continue the data analysis considering these covariates in their natural scale. So, we could easily interpret their effects on tree survival.

Bayesian information criterion (BIC) and residual deviance (RD) indicated that the complementary log-log link function provided the best fit in both modeling approaches (TABLE 4.1). However, when BIC values were compared between covariate selection approaches, the stepwise procedure provided the best fit for all link functions. This result is related to the largest penalty on the log-likelihood function of the model based on the regularization approach due to the largest number of parameters. Furthermore, complement log-log and probit link functions had the same covariates selected by the stepwise procedure. For logit link function, this method also selected the covariates related to thinnings, such as *gthin* and *nthin*, while Cauchit selected *ddom* and *daverage*.

We performed a graphical analysis for evaluating the assumptions of the fitted models. Our models were based on a Bernoulli specification of the binary response variable *survival*. Thus, the assumptions usually assumed for normal data are no longer demanded. The half-Normal plot presented in FIGURE 4.3 suggested that the models were properly specified, once the residuals do not exceed the simulated envelope. However, both models presented similar behavior, indicating a good fit and a suitable probability distribution of response variable. As a feature of the randomly quantile residuals, when the model is suitable to the data it should be expected a normal distribution of the residuals, regardless of the distribution of the response variable and selected covariates. In our case, sample and theoretical residual quantile had a linear association (FIGURE 4.3), confirming a good performance of the fitted models and a normal distribution of the residuals.

Some preliminary analyses indicated that only the main effects of covariates were suitable for modeling the response variable *survival,* in which interaction terms are not required to be included in the linear predictor. Parameter estimates and standard errors for both covariate selection procedures are presented in TABLE 4.2. Point estimates of the fitted model based on stepwise selection suggested that the response variable has negative relation to *gsample* and *dcv*, since the associated parameters had a negative sign, as can be observed in FIGURE 4.4. In practice, larger cross-sectional area and higher diameter variability in the sample are associated with a lower individual survival probability. On the other side, *nsample*, *dg*, and *dmax* covariates are associated with higher values of survival probability (FIGURE 4.4).

TABLE 4.1 - BAYESIAN INFORMATION CRITERION (BIC) AND RESIDUAL DEVIANCE (RD) BY LINK FUNCTIONS AND COVARIATE SELECTION METHODS

Link function	BIC (Number of variables)		Residual deviance	
	Stepwise	Regularization	Stepwise	Regularization
Cauchit	7,068.31 (9)	7,094.78 (12)	6,963.30	6,958.20
C. log-log	6,847.46 (5)	6,904.20 (12)	6,784.40	6,768.30
Logit	6,874.73 (7)	6,910.75 (12)	6,790.70	6,774.20
Probit	6,851.50 (5)	6,906.84 (12)	6,788.50	6,769.30

TABLE 4.2 - PARAMETER ESTIMATES, STANDARD ERRORS (SE) AND P-VALUE OF THE FITTED MODELS WITH COMPLEMENT LOG-LOG LINK FUNCTION ON THE LINEAR PREDICTOR SCALE

Parameter	Estimate	SE	p-value	Estimate	SE	p-value
	Regularization			Stepwise		
intercept	-0.2940	0.3917	$p > 0.05$	-0.3973	0.2305	$p \leq 0.10$
age	-0.0097	0.0128	$p > 0.05$	-	-	-
gsample	-0.0404	0.0034	$p \leq 0.05$	-0.0413	0.0024	$p \leq 0.05$
nsample	0.0017	0.0001	$p \leq 0.05$	0.0017	0.0001	$p \leq 0.05$
daverage	-0.4005	0.3486	$p > 0.05$	-	-	-
dcv	-0.0484	0.0146	$p \leq 0.05$	-0.0411	0.0047	$p \leq 0.05$
dg	0.4891	0.3469	$p > 0.05$	0.0575	0.0118	$p \leq 0.05$
dmax	0.0451	0.0093	$p \leq 0.05$	0.0312	0.0069	$p \leq 0.05$
ddom	-0.0426	0.0224	$p > 0.05$	-	-	-
hdom	0.0028	0.0102	$p > 0.05$	-	-	-
thinsample	-0.0238	0.1884	$p > 0.05$	-	-	-
gthin	0.0230	0.0098	$p \leq 0.05$	-	-	-
nthin	-0.0008	0.0003	$p \leq 0.05$	-	-	-

FIGURE 4.3 - HALF-NORMAL PLOT (LEFT) AND RANDOMLY QUANTILE RESIDUALS (RIGHT) FOR DIAGNOSING THE FITTED MODELS

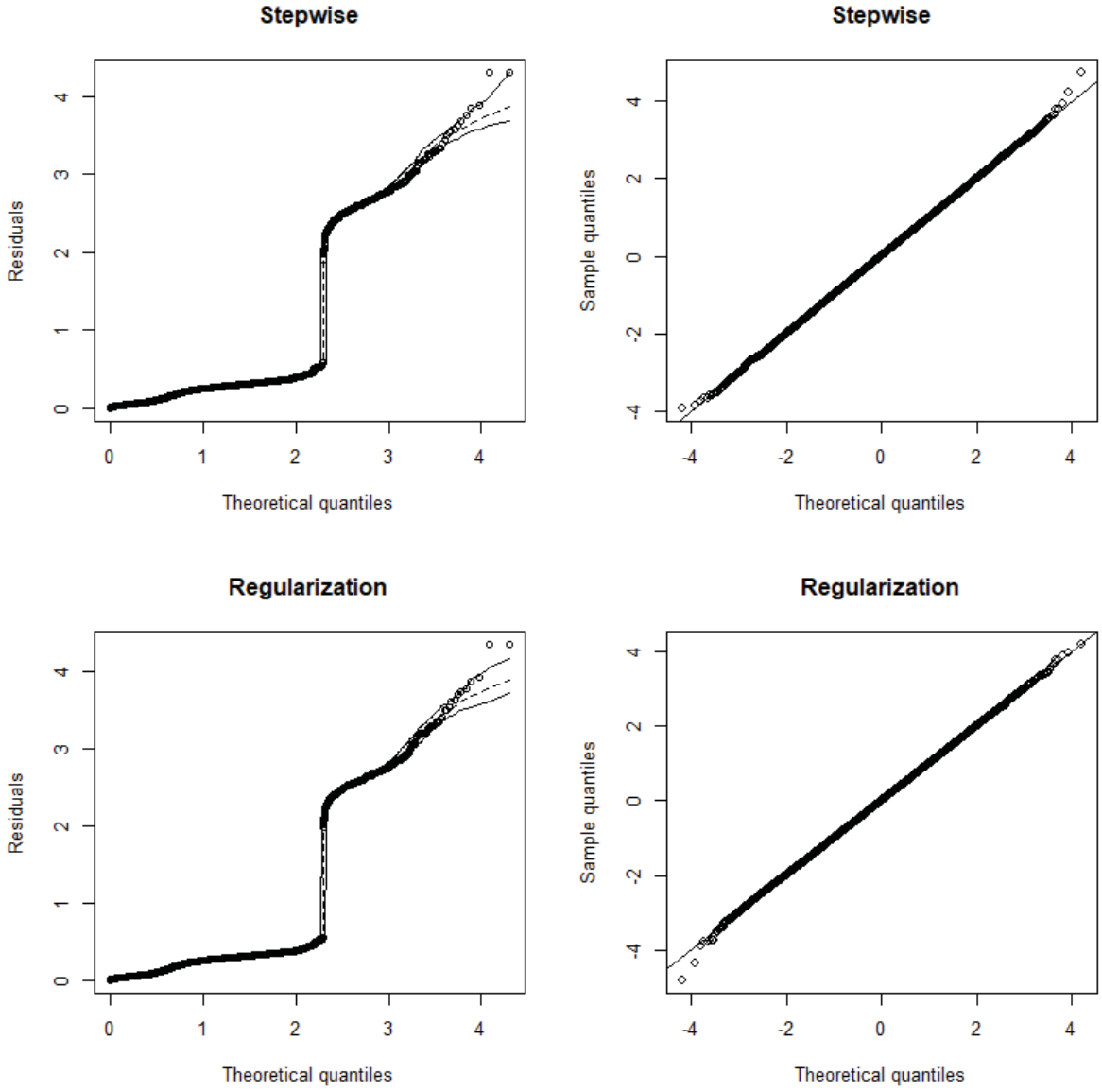
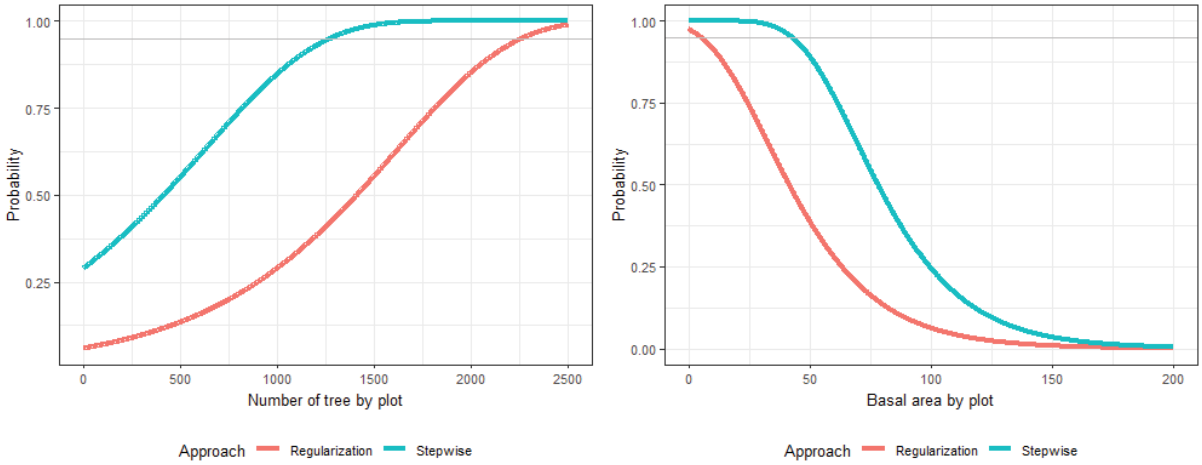


FIGURE 4.4 - PREDICTION OF SURVIVAL PROBABILITY FOR THE COVARIATES *nsample* AND *gsample* OF THE FITTED MODELS. COVARIATES WERE FIXED AT THE MEAN VALUES.



4.3.3. PREDICTIVITY PERFORMANCE

A validation data set was used for comparing the performance of the fitted models in predicting the response variable, once the forest planning directly depends on the estimated number of alive trees in a forest stand. The ROC curves were similar for both models (FIGURE 4.5). However, the area under the curve was 0.805 for the model selected by stepwise procedure, and 0.814 for the model chosen by regularization method, indicating a slightly better predictions for the model with more parameters.

When we changed the cut point for defining a suitable probability value for classifying trees in survivors or non-survivors, the best result was obtained with a 0.95 probability cut point. This result was observed for both models, once that in this probability cut point was obtained the highest value in Younden´s rule and the lowest value in Closest Topleft´s rule (TABLE 4.3). We also noticed that the model based on the regularization procedure presented slightly higher values in the decision rules than the stepwise procedure, resulting in a better performance for classifying the individuals.

The estimated sensitivity and specificity values for a 0.95 probability cut point are presented in Table 4.4. The results suggested that the models have great capacity to identify alive trees, due to the high sensitivity value; while low values were observed for specificity, which implies in difficult to identify dead trees. Still, negative predictive value (NPV) suggested that the probability of a tree to be dead given that the model predicted as a non-

survival tree was about 0.08 for both models, which is directly related to the low incidence of mortality in forest stands. However, the probability of a tree to be alive given that the model predicted as a survival tree was 0.99, as observed on the positive predictive value (PPV).

FIGURE 4.5 - ROC CURVE OF THE MODELS APPLIED TO THE VALIDATION DATA SET

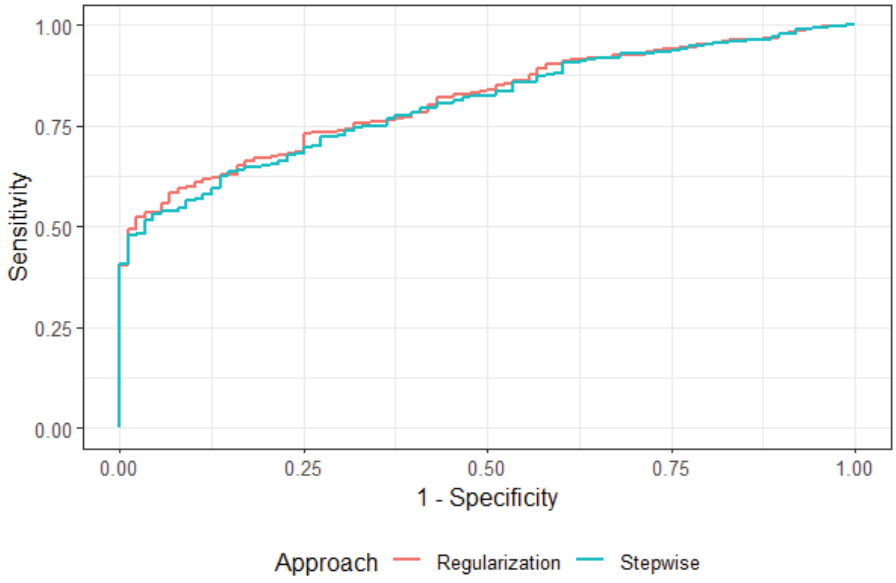


TABLE 4.3 - YOUDEN AND CLOSEST TOPLEFT DECISION RULES FOR DIFFERENT PROBABILITY CUT POINTS OF THE MODELS APPLIED TO THE VALIDATION DATA SET

Model	Cut point	Youden	Closest Topleft
Stepwise	0.75	1.000	1.000
	0.85	1.010	0.977
	0.90	1.049	0.890
	0.95	1.300	0.372
Regularization	0.75	1.000	1.000
	0.85	1.000	1.000
	0.90	1.049	0.890
	0.95	1.315	0.347

TABLE 4.4 – SENSITIVITY, SPECIFICITY, POSITIVE PREDICTIVE VALUE (PPV) AND NEGATIVE PREDICTIVE VALUE (NPP) BY SELECTED MODELS APPLIED TO THE VALIDATION DATA SET FOR 0.95 PROBABILITY CUT POINT

Model	Sensitivity	Specificity	PPV	NPP
Stepwise	0.902	0.398	0.985	0.084
Regularization	0.895	0.420	0.986	0.083

4.4. DISCUSSION

The main goal of this paper was to specify and fit a generalized linear model for estimating the tree survival probability in loblolly pine plantations. We tested four strategies of covariate selections based on stepwise and regularization procedures, such as ridge regression, lasso and elastic net method. We were also interested in analyzing the influence of link functions when selecting covariates for composing the linear predictor. Initially, we expected that the regularization method would be more appropriate for selecting correlated covariates, which are common in forest variables, because this approach can include some bias in parameter estimates in contrast to reduce their variance. Since the covariates are correlated and standard errors are larger, regularization procedures are quite promising in forest modeling. However, the penalization term had no effect in our model. As consequence, the stepwise procedure performed best due to the fewer selected covariates, making this a more parsimoniously procedure.

A different number of selected covariates for composing the linear predictor of the model can be obtained when we consider different link functions, what suggests that the link function must be appropriated for a specified data set. Despite preference by Logit link function on the tree survival probability modeling in forest plantations (TÉO, 2017; YANG et al., 2003; YAO et al., 2001), better results on BIC were obtained for complementary log-log and probit link functions, which provided models with a few parameters. The performance of the complementary log-log link function showed evidence that the behavior of tree survival probability is asymmetric when related to the linear predictor, once that the individual tree survival probability approaches to zero and one in different rate. Thus, considering a symmetric link function may not be a reasonable assumption in tree survival modeling (JIANG et al., 2013). These results became relevant because the probability of success presented values quite near of one, where the link functions show more discrepancy.

Our models performed well for fitting and predicting the survival probabilities. However, better results can be obtained whether more covariates are considered for composing the linear predictor, such as environmental variables, mainly whether the model is applied to large areas. Zhang et al. (2017) modeled the mortality of forest plantations located at China using climatic covariates, besides initial planting density and competitions indexes. The authors suggested the inclusion of climatic variables in mortality models can facilitate the projection of

tree mortality under future climate change conditions. Thapa & Burkhart (2015) tested climatic and soil effects on tree mortality, and the predictions performed best when they included these covariates. However, climatic variables were significant just when the model was fitted for large areas, which suggests that only climatic effects play a minor role in small areas.

In forest research involving tree mortality or survival, tree competition indices are commonly used as predictor variables (MIRANDA et al., 2017; TÉO, 2017, ZHANG et al., 2017). However, these indexes are computed in function of covariates usually included in the linear predictor. As an example, basal area larger index (BAL) is obtained by summing the cross-sectional area of all trees with larger diameter than the object tree (EID & TUHUS, 2001), then being a function of the diameter at breast height. This procedure can induce a correlation between both variables (MIRANDA, 2016; SCHRÖDER & GADOW, 1999). Consequence of correlated covariates is a larger standard error of the parameters estimates, that can compromise the hypothesis tests and inferences. In our preliminary analysis, changes in the parameters sign and standard error of the stepwise model were observed when we removed the covariates *dg* or *dmax*. This result is explained by the high correlation value (0.95) between them.

We tested *thinsample*, *gthin* and *nthin* covariates for accounting possible thinning effects on tree survival probability. However, similar to what was found by Avila & Burkhart (1992), no improvement was obtained in the predictions when those variables were added to the model. A possible reason is that the mortality is a quite rare phenomenon, and after thinning we also do not expect a relevant regular mortality. According to Bose et al. (2018), commercial thinning treatment replaced self-thinning of suppressed trees; thus, decreasing tree mortality in loblolly pine and Douglas-fir plantations in North America. The authors also highlighted that the thinning was effective for reducing long-term tree mortality in red spruce and balsam fir, confirming the significance of thinning intensity and basal area as relevant predictor covariates.

Our tree survival probability models presented a great ability to predict alive trees, as suggested by the sensitivity statistic (Table 5). Téo (2017) used logistic regression combined with logit link function for modeling *Pinus taeda* tree survival probability in Midwest of Santa Catarina. The sensitivity of his model was 98.9% and the specificity was 43.1% for irregular mortality, being similar that one obtained in this paper. When the author considered only regular mortality, sensitivity and specificity were 99.1% and 52.3%, respectively. These results suggest that the natural mortality is more regular than that one caused by external factors.

A possible reason for the discrepancy observed for sensitivity and specificity of the model was the different number of survived and non-survived trees. These imbalance between alive and dead trees classes have influence on the effectiveness of the model. In general, tree

survival probability models usually do not present high values of specificity (ADAME et al., 2010; TÉO, 2017), what may be related to the few dead trees in a forest plantation when compared to the number of alive trees. As alternative, Kuhn & Johnson (2013) suggested that a balanced training set may help to deal with this class imbalance. However, this approach still requires detailed researches in forest modeling. Another possible reason for lower values of specificity is related to the lack of ability of the covariates usually measured at forest inventories in identifying the dead individuals. Thus, we recommend testing more covariates for increase the specificity of the survival models.

Finally, futures topics to be explored in survival modeling are related to the inclusion of forest inventories performed in several occasions, with several occasions of sample units measurements, defining a longitudinal study, due to the temporal dependency among observations. Applications of spatial statistics should be considered for improving the analysis of tree survival in forest stands, since the environmental gradient can influence the tree individual mortality.

4.5. CONCLUSION

In this study, we specify and fit a generalized linear model for estimating the probability of loblolly pine tree survival in forest plantations, considering covariates usually measured in forest inventories. The plot-level variables that most contributed to assess tree survival were basal area, number of trees, maximum diameter, diameter of the average cross-sectional area and the diameter coefficient of variation.

The stepwise procedure for selecting covariates was more parsimonious than the regularization procedures tested; and combined with complementary log-log link function was the procedure provided the most suitable model. The model presented a great prediction ability, mainly due to the high number of survival trees. Additional researches related to regularization techniques are recommended in forest modeling, mainly regarding survival and individual growth models.

REFERENCES

ADAME, P.; RÍO, M. DEL; CAÑELLAS, I. Modeling individual-tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. **Annals of Forest Science**, v. 67, n. 8, 2010.

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L.M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711-728, 2013.

AVILA, O.B.; BURKHART, H.E. Modeling survival of loblolly pine trees in thinned and unthinned plantations. **Canadian Journal of Forest Research**, v. 22, n. 12, p. 1878-1882, 1992.

BOSE, A.K.; WEISKITTEL, A.; KUEHNE, C.; WAGNER, R.G.; TURNBLOM, R.; BURKHART H.E. Tree-level growth and survival following commercial thinning of four major softwood species in North America. **Forest Ecology and Management**, v. 427, p. 355-364, 2018.

DAS, A.J.; STEPHENSON, N.L. Improving estimates of tree mortality probability using potential growth rate. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 920-928, 2015.

DÍEGUEZ-ARANDA, U.; CASTEDO-DORADO F.; ÁLVAREZ-GONZÁLEZ J.G.; RODRÍGUEZ-SOALLEIRO. Modeling mortality of Scot Pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. **European Journal of Forest Research**, v. 124, p. 143-153, 2005.

DUNN, P.K.; SMYTH G.K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 236-244, 1996.

EID, T.; TUHUS, E. Model for individual tree mortality in Norway. **Forest Ecology and Management**, v. 154, p. 69-84. 2001.

FRIEDMAN J.; HASTIE, T.; TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1-22, 2010.

GIOLO, S.R. **Introdução à análise de dados categóricos com aplicações**. Editora: Blucher – Projeto Fisher ABE, 2017.

IBÁ. **Industria brasileira de árvores**. Available in http://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf. 2017.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Manuais técnicos em geociências: Manual técnico da vegetação brasileira**, n. 1, 2° ed. Rio de Janeiro: IBGE, 2012. 275 p.

JIANG, B.X.; DEY, D.K.; PRUNIER, R.; WILSON, A.M.; HOLSINGER, K.E. A new class of flexible link functions with applications to species co-occurrence in Cape florist region. **The Annals of Applied Statistics**, v. 7, n. 4, p. 2180-2204, 2013.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer, 2013. 600 p.

MACFARLANE, D.W.; WEISKITTEL, A.R. A new method for capturing stem taper variation for trees of diverse morphological types. **Canadian Journal of Forest Research**, v. 46, n. 6, p. 804–815, 2016.

MCCULLAGH, P.; NELDER, J.A. **Generalized Linear Models**. Chapman & Hall, 2° ed., 1989.

MEHTÄTALO, L.; DE-MIGUEL, S.; GREGOIRE, T.G. Modeling height-diameter curves for prediction. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 826–837, 2015.

MIRANDA, R.O.V. **Modelagem de árvores individuais para povoamentos não desbastados de *Pinus taeda* L.** 169 f. Doutorado (Doutorado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2016.

MIRANDA R.O.V.; FIGUEIREDO FILHO, A.; MACHADO, S.A.; CASTRO, R.V.O.; FIORENTIN, L.D.; BERNETT, L.G. Modelagem da mortalidade em povoamentos de *Pinus taeda* L. **Scientia Forestalis**, v.15, n.115, p.435-444, 2017.

MINGOTI, S.L. **Análise de Dados Através de Métodos de Estatística Multivariada**. Belo Horizonte: Editora UFMG, 2005.

MORAL, R.A.; HINDLE, J.; DEMÉTRIO, C.G.B. Half-normal plots and overdispersed models in R: The hnp package. **Journal of Statistical Software**, v. 81, 2017.

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2019.

SABATIA, C.O.; BURKHART, H.E. On the use of upper stem diameters to localize a segmented taper equation to new trees. **Forest Science**, v. 61, n. 3, p. 411–423, 2015.

SCHÖDER, J.; VON GADOW, K. Testing a new competition index for Maritime pine in northwestern Spain. **Canadian Journal of Forest Research**, v. 29, n. 2, p. 280-283, 1999.

SING T.; SANDER O.; BEERENWINKEL N.; LENGAUER T. ROCR: visualizing classifier performance in R. **Bioinformatics**, v. 21, n. 20, p. 7881, 2005.

SZMYT, J.; TARASIUK, S. Species-specific spatial structure, species coexistence and mortality pattern in natural, uneven-aged Scots pine (*Pinus sylvestris* L.) - dominated forest. **European Journal of Forest Research**, v. 137, n. 1, p. 1–16, 2018.

TÉO, S.J. **Modelagem do crescimento e produção de árvore individual independente da distância, para *Pinus taeda* L., na região meio oeste do estado de Santa Catarina**. 2017. 283 p. Doutorado (Doutorado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267-288, 1996.

THAPA, R.; BURKHART, H.E. Modeling stand-level mortality of loblolly pine (*Pinus taeda* L.) using stand, climate, and soil variables. **Forest Science**, v. 61, n. 5, p. 834-846, 2015.

TRINCADO, G.; VANDERSCHAAF, C.L.; BURKHART, H.E. Regional mixed-effects height-diameter models for loblolly pine (*Pinus taeda* L.) plantations. **European Journal of Forest Research**, v. 126, n. 2, p. 253–262, 2007.

UNAL, I. Defining an optimal cut-point value in ROC: analysis: an alternative approach. **Computational and Mathematics Methods in Medicine**, 2017.

VANCLAY, J.K. Mortality functions for North Queensland rain forests. **Journal of Tropical Forest Science**, v. 4, n. 1, p. 15-36, 1991.

WESTFALL, J.A.; SCOTT, C.T. Taper models for commercial tree species in the northeastern United States. **Forest Science**, v. 56, n. 6, p. 515–528, 2010.

YANG, Y.; TITUS, S.J.; HUANG, S. Modeling individual tree mortality for white spruce in Alberta. **Ecological Modelling**, v. 163, n. 3, p. 209-222, 2003.

YAO, X.; TITUS, S.J.; MACDONALD, S.E. A generalized logistic model of individual tree mortality of aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. **Canadian Journal of Forest Research**, v. 31, n. 2, p. 283-291, 2001.

ZHANG, X.; CAO, Q.V.; DUAN, A.; ZHANG J. Modeling tree mortality in relation to climate, initial planting density, and competition in Chinese fir plantations using a Bayesian logistic multilevel method. **Canadian Journal of Forest Research**, v. 47, p. 1278-1285, 2017.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal Royal of Statistical Society. Series B**, v. 67, n. 2, p. 301-320, 2005.

5. GENERAL CONCLUSIONS AND RECOMMENDATIONS

We introduced new statistical methods with potential to be applied to forest biometrics in this thesis. In general, the methods presented great performance to handle with common problems in forest modeling. We highlighted the ability of the models for handling in a relatively easy way with non-independent observations and the correlated forest variables.

The covariance generalized linear models was suitable for *Pinus taeda* stem taper modeling. This class of models increased the knowledge about the behavior of stem taper and how the correlation patter is over the tree stem. The response variable predictions were improved when conditionate predictions were performed for relative diameters.

The multivariate covariance generalized linear models showed that a jointly modeling is recommended for estimating the *Araucaria angustifolia* height and volume, due to the correlation between them. Variance functions must be used for explaining the relationship between mean and variance of variable volume. Regular behavior was observed for the variable height and the variance function was not required.

The univariate and multivariate cases of covariance generalized linear models has large applications to forest biometrics. The uncertainties related to the response variable are easily quantified in confidence intervals; and a more robust predictions can be performed due to the covariance matrix. Additional research is recommended to biomass and carbon modeling using the methodologies presented in this thesis. We also recommended additional studies involving linear and non-linear models for growth-yield systems at individual and non-individual tree level.

The Bernoulli generalized linear model presented a great prediction ability. The covariate selection based on penalization procedures require more research. Applications of this methodology for covariate selection is also recommended for growth-yield modeling, due to the high correlation among independent variables.

REFERENCES

ADAME, P.; RÍO, M. DEL; CAÑELLAS, I. Modeling individual-tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. **Annals of Forest Science**, v. 67, n. 8, 2010.

ARIAS-RODIL, M.; CASTEDO-DORADO, F.; CÁMARA-OBREGÓN, A.; DIÉGUEZ-ARANDA, U. Fitting and calibrating a multilevel mixed-effects stem taper model for Maritime Pine in NW Spain. **PlosOne**, v. 10, 2015a.

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L.M.; SPAROVEK, G. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711-728, 2013.

AVILA, O.B.; BURKHART, H.E. Modeling survival of loblolly pine trees in thinned and unthinned plantations. **Canadian Journal of Forest Research**, v. 22, n. 12, p. 1878-1882, 1992.

ARIAS-RODIL, M.; DIÉGUEZ-ARANDA, U.; PUERTA, F.R.; LÓPEZ-SÁNCHEZ, C.A.; LÍBANO, E.C.; OBREGÓN, A.C.; CASTEDO-DORADO. Modeling and localizing a stem taper function for *Pinus radiata* in Spain. **Canadian Journal of Forest Research** , v. 45, p. 647–658, 2015b.

ARIAS-RODIL, M.; DIÉGUEZ-ARANDA, U.; BURKHART, H.E. Effects of measurement error in total tree height and upper-stem diameter on stem volume prediction. **Forest Science**, v. 63, n. 3, p. 250–260, 2017.

BERGER, A.; GSCHWANTNER, T.; MACROBERTS, R.E.; SCHADAUER K. Effects of measurement errors on individual tree stem volume estimates for the Austrian National Forest Inventory. **Forest Science**, v. 60, n. 1, p. 14–24, 2014.

BONAT, W.H. Modelling mixed types of outcomes in additive genetic models. **The international journal of biostatistics**, v. 13, n. 2, p. 1–16, 2017.

BONAT, W.H. Multiple response variables regression models in R: The mcglm Package. **Journal of Statistical Software**, v. 84, n. 4, 2018.

BONAT, W.H.; JØRGENSEN, B. Multivariate covariance generalized linear models. **Journal of the Royal Statistical Society. Series C: Applied Statistics**, v. 65, n. 5, p. 649–675, 2016.

BONAT, W.H.; OLIVERO, J.; GRANDE-VEJA, M.; FARFÁN A.; FA, J.E. Modelling the Covariance Structure in Marginal Multivariate Count Models: Hunting in Bioko Island. **Journal of Agricultural, Biological, and Environmental Statistics**, v. 22, n. 4, p. 446–464, 2017.

BOSE, A.K.; WEISKITTEL, A.; KUEHNE, C.; WAGNER, R.G.; TURNBLOM, E.; BURKHART, H.E. Tree-level growth and survival following commercial thinning of four major softwood species in North America. **Forest Ecology and Management**, v. 427, p. 355–364, 2018.

BURKHART, H.E.; TOMÉ, M. **Modeling forest trees and stands**. New York: Springer, 2012. 457 p.

CAO, Q.V.; WANG, J. Evaluation of methods for calibrating a tree taper equation. **Forest Science**, v. 61, n. 2, p. 213–219, 2015.

CASTEDO-DORADO, F.; DIÉGUEZ-ARANDA, U.; ANTA, M.B.; RODRÍGUEZ, M.S.; VON GADOW, K. A generalized height-diameter model including random components for radiata pine plantations in northwestern Spain. **Forest Ecology and Management**, v. 229, n. 3, p. 202–213, 2006.

DAS, A.J.; STEPHENSON, N.L. Improving estimates of tree mortality probability using potential growth rate. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 920-928, 2015.

DE-MIGUEL, S; MEHTÄTALO, L.; SHATER, Z.; KRAID, B.; PUKKALA, T. Evaluating marginal and conditional predictions of taper models in the absence of calibration data. **Canadian Journal of Forest Research**, v. 42, n. 7, p. 1383–1394, 2012.

DIÉGUEZ-ARANDA, U., CASTEDO-DORADO, F.; ÁLVAREZ-GONZÁLEZ, J.G.; ROJO, A. Compatible taper function for Scots pine plantations in northwestern Spain. **Canadian Journal of Forest Research**, v. 36, n. 5, p. 1190–1205, 2006.

DUNN, P.K.; SMYTH G.K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, n. 3, p. 236-244, 1996.

EID, T.; TUHUS, E. Model for individual tree mortality in Norway. **Forest Ecology and Management**, v. 154, p. 69-84. 2001.

FIORENTIN, L.D.; BONAT, W.H.; PELISSARI, A.L.; MACHADO, S.A.; TÉO, S.J. Modelagem marginal conjunta da altura e volume para *Araucaria angustifolia*. **Biofix Scientific Journal**, v. 5, n. 1, p. 121–129, 2020.

FORTIN, M.; ROBERT, N.; MANSO, R. Uncertainty assessment of large-scale forest growth predictions based on a transition-matrix model in Catalonia. **Annals of Forest Science**, v. 73, n. 4, p. 871–883, 2016.

FORTIN, M.; SCHNEIDER, R.; SAUCIER, J. Volume and error variance estimation using integrated stem taper models. **Forest Science**, v. 59, n. 3, 2013.

FRIEDMAN J.; HASTIE, T.; TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1-22, 2010.

FU, L.; ZHANG, H,; SHARMA, R.P.; PANG, L.; WANG, G. A generalized nonlinear mixed-effects height to crown base model for Mongolian oak in northeast China. **Forest Ecology and Management**, v. 384, p. 34–43, 2017.

GIOLO, S.R. **Introdução à análise de dados categóricos com aplicações**. Editora: Blucher – Projeto Fisher ABE, 2017.

GOMAT, H.Y.; DELEPORTE, P.; MOUKINI, R.; MIALOUNGUILA, G.; OGNOUABI, N.; SAYA, A.R.; VIGNERON, P.; SAINT-ANDRE, L. What factors influence the stem taper of Eucalyptus: Growth, environmental conditions, or genetics? **Annals of Forest Science**, v. 68, n. 1, p. 109–120, 2011.

GÓMEZ-GARCÍA, E.; CRECENTE-CAMPO, F.; DIÉGUEZ-ARANDA, U. Selection of mixed-effects parameters in a variable-exponent taper equation for birch trees in northwestern Spain. **Annals of Forest Science**, v. 70, n. 7, p. 707–715, 2013.

IBÁ. **Industria brasileira de árvores**. Available in http://iba.org/images/shared/Biblioteca/IBA_RelatorioAnual2017.pdf. 2017.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Manuais técnicos em geociências: Manual técnico da vegetação brasileira**, n. 1, 2° ed. Rio de Janeiro: IBGE, 2012. 275 p.

JIANG, B.X.; DEY, D.K.; PRUNIER, R.; WILSON, A.M.; HOLSINGER, K.E. A new class of flexible link functions with applications to species co-occurrence in Cape florist region. **The Annals of Applied Statistics**, v. 7, n. 4, p. 2180-2204, 2013.

KOZAK, A. Effects of multicollinearity and autocorrelation on the variable-exponent taper functions. **Canadian Journal of Forest Research**, v. 27, n. 5, p. 619–529, 1997.

KOZAK, A. My last words on taper equations. **Forestry Chronicle**, v. 80, n. 4, p. 507–515, 2004.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. Springer, 2013. 600 p.

KUBLIN, E.; BREIDENBACH, J.; KÄNDLER, G.A. flexible stem taper and volume prediction method based on mixed-effects B-spline regression. **European Journal of Forest Research**, v. 132, n. 5–6, p. 983–997, 2013.

LAM, T.Y.; KERSHAW, J.A.; HAJAR, Z.S.N.; RAHMAN, K.A.; WEISKITTEL, A.R.; POTTS, M.D. Evaluating and modelling genus and species variation in height-to-diameter relationships for Tropical Hill Forests in Peninsular Malaysia. **Forestry**, v. 90, n. 2, p. 268–278, 2017.

LAPPI, J.A. multivariate, nonparametric stem-curve prediction method. **Canadian Journal of Forest Research**, v. 36, n. 4, p. 1017–1027, 2006.

LEJEUNE, G., UNG, C.H.; FORTIN, M.; GUO, X.J.; LAMBERT, M.C.; RUEL, J.C. A simple stem taper model with mixed effects for boreal black spruce. **European Journal of Forest Research**, v. 128, n. 5, p. 505–513, 2009.

LEE, Y.; NELDER, J.A. Conditional and marginal models: Another view. **Statistical Science**, v. 19, n. 2, p. 219–238, 2004.

LI, R.; WEISKITTEL, A. Development and evaluation of regional taper and volume equations for the primary conifer species in the Acadian Region of North America. **Annals of Forest Science**, v. 67, p. 21–24, 2010.

MACFARLANE, D.W.; WEISKITTEL, A.R. A new method for capturing stem taper variation for trees of diverse morphological types. **Canadian Journal of Forest Research**, v. 46, n. 6, p. 804–815, 2016.

MCCULLAGH, P.; NELDER, J.A. **Generalized Linear Models**. Chapman & Hall, 2° ed., 1989.

MACHADO, S.A.; FIGUEIREDO FILHO, A. **Dendrometria**. Guarapuava: UNICENTRO, 2009.

MARCHIORI, J.N.C. **Dendrologia das gimnospermas.** Editora ufsm, ed. 2, 2005.

MÄKINEN, H.; JYSKE, T.; NÖJD, P. Dynamics of diameter and height increment of Norway spruce and Scots pine in southern Finland. **Annals of Forest Science**, v. 75, n. 1, p. 1–11, 2018.

MACPHEE, C.; KERSHAW, J.A.; WEISKITTEL, A.R.; GOLDING, J.; LAVIGNE, M.B. Comparison of approaches for estimating individual tree height-diameter relationships in the Acadian forest region. **Forestry: An international Journal of Forest Research**, v. 91, n. 1, p. 132–146, 2018.

MANSO, R.; NINGRE, F.; FORTIN, M. Simultaneous prediction of plot-level and tree-level harvest occurrences with correlated random effects. **Forest Science**, v. 64, n. 5, p. 461–470, 2018.

MAX, T.; BURKHART, H. Segmented polynomial regression applied to taper equations. **Forest Science**, v. 22, n. 3, p. 283–289, 1976.

MCROBERTS, R.E.; WESTFALL, J,A. Effects of uncertainty in model predictions of individual tree volume on large area volume estimates. **Forest Science**, v. 60, n. 1, p. 34–42, 2014.

MEHTÄTALO, L.; DE-MIGUEL, S.; GREGOIRE, T.G. Modeling height-diameter curves for prediction. **Canadian Journal of Forest Research**, v. 45, n. 7, p. 826–837, 2015.

MIRANDA, R.O.V. **Modelagem de árvores individuais para povoamentos não desbastados de *Pinus taeda* L.** 169 f. Doutorado (Doutorado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba, 2016.

MIRANDA R.O.V.; FIGUEIREDO FILHO, A.; MACHADO, S.A.; CASTRO, R.V.O.; FIORENTIN, L.D.; BERNETT, L.G. Modelagem da mortalidade em povoamentos de *Pinus taeda* L. **Scientia Forestalis**, v.15, n.115, p.435-444, 2017.

MINGOTI, S.L. **Análise de Dados Através de Métodos de Estatística Multivariada**. Belo Horizonte: Editora UFMG, 2005.

MORAL, R.A.; HINDLE, J.; DEMÉTRIO, C.G.B. Half-normal plots and overdispersed models in R: The hnp package. **Journal of Statistical Software**, v. 81, 2017.

NASCIMENTO, R.G.M.; MACHADO, S.A.; FIGUEIREDO FILHO, A.; HIGUCHI, N. A

growth and yield projection system for a tropical rainforest in the Central Amazon, Brazil. **Forest Ecology and Management**, v. 327, p. 201–208, 2014.

OIJEN, V.M. Bayesian methods for quantifying and reducing uncertainty and error in forest models. **Current Forestry Reports**, v. 3, n. 4, p. 269–280, 2017.

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2019.

RIOFRÍO, J.; DEL RÍO, M.; BRAVO, F. Mixing effects on growth efficiency in mixed pine forests. **Forestry: An international Journal of Forest Research**, v. 90, n. 3, p. 381–392, 2017.

SABATIA, C.O. Use of upper stem diameters in a polynomial taper equation for New Zealand radiata pine: an evaluation. **New Zealand Journal of Forestry Science**, v. 46, n. 1, 2016.

SABATIA, C.O.; BURKHART, H.E. On the Use of Upper Stem Diameters to Localize a Segmented Taper Equation to New Trees. **Forest Science**, v. 61, n. 3, 2015.

SCHEEREN, L.W.; FINGER, C.A.G.; SCHUMACHER, M.V.; LONGHI, S.L. Crescimento em altura de *Araucaria angustifolia* (Bert.) O. Ktze. em três sítios naturais, na região de Canela (RS). **Ciência Florestal**, v. 9, n. 2, p. 23–40, 1999.

SCHÖDER, J.; VON GADOW, K. Testing a new competition index for Maritime pine in northwestern Spain. **Canadian Journal of Forest Research**, v. 29, n. 2, p. 280-283, 1999.

SING T.; SANDER O.; BEERENWINKEL N.; LENGAUER T. ROCR: visualizing classifier performance in R. **Bioinformatics**, v. 21, n. 20, p. 7881, 2005.

SEKI, M.; SAKICI, O.E. Dominant height growth and dynamic site index models for crimean pine in the Kastamonu-Taşköprü region of Turkey. **Canadian Journal of Forest Research**, v. 47, n. 11, p. 1441–1449, 2017.

SHARMA, M.; REID, D.E.B. Stand height/site index equations for jack pine and black spruce trees grown in natural stands. **Forest Science**, v. 64, n. February, p. 33–40, 2017.

SHARMA, R. P.; VACEK, Z.; VACEK, S.; JANSA, V.; KUCERA, M. Modelling individual tree diameter growth for Norway spruce in the Czech Republic using a generalized algebraic difference approach. **Journal of Forest Science**, v. 63, n. 5, p. 227–238, 2017.

STOKLOSA, J.; GIBB, H.; WARTON, D.I. Fast forward selection for generalized estimating equations with a large number of predictor variables. **Biometrics**, v. 70, n. 1, p. 110–120, 2014.

SZMYT, J.; TARASIUK, S. Species-specific spatial structure, species coexistence and mortality pattern in natural, uneven-aged Scots pine (*Pinus sylvestris* L.) - dominated forest. **European Journal of Forest Research**, v. 137, n. 1, p. 1–16, 2018.

TENZIN, J.; TENZIN, K.; HASENAUER, H. Individual tree basal area increment models for broadleaved forests in Bhutan. **Forestry: An international Journal of Forest Research**, v. 90, n. 3, p. 367–380, 2017.

TÉO, S.J. **Modelagem do crescimento e produção de árvore individual independente da distância, para *Pinus taeda* L., na região meio oeste do estado de Santa Catarina**. 2017. 283 p. Doutorado (Doutorado em Engenharia Florestal) – Universidade Federal do Paraná, Curitiba.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. **Journal of the Royal Statistical Society**, v. 58, n. 1, p. 267-288, 1996.

THAPA, R.; BURKHART, H.E. Modeling stand-level mortality of loblolly pine (*Pinus taeda* L.) using stand, climate, and soil variables. **Forest Science**, v. 61, n. 5, p. 834-846, 2015.

THOMAS C.; ANDRADE, C.M.; SCHNEIDER, P.R.; FINGER, C.A.G. Comparação de equações volumétricas ajustadas com dados de cubagem e análise de tronco. **Ciência Florestal**, v. 16, n. 3, p. 319–327, 2006.

TRINCADO, G.; VANDERSCHAAF, C.L.; BURKHART, H.E. Regional mixed-effects height-diameter models for loblolly pine (*Pinus taeda* L.) plantations. **European Journal of Forest Research**, v. 126, n. 2, p. 253–262, 2007.

UNAL, I. Defining an optimal cut-point value in ROC: analysis: an alternative approach. **Computational and Mathematics Methods in Medicine**, 2017.

VANCLAY, J.K. Mortality functions for North Queensland rain forests. **Journal of Tropical Forest Science**, v. 4, n. 1, p. 15-36, 1991.

WEBER, K.S.; SANQUETA, C.R.; EISFELD, R.L. Variação volumétrica e distribuição espacial do estoque de carbono Em Floresta Ombrófila Mista. **Revista Acadêmica: Ciência Animal**, v. 3, n. 2, p. 77–85, 2017.

WESTFALL, J.A.; SCOTT, C.T. Taper models for commercial tree species in the northeastern United States. **Forest Science**, v. 56, n. 6, p. 515–528, 2010.

WICKHAM, H. **Ggplot 2: elegant graphics for data analysis**. Springer, 2016.

VERBEKE, G.; FIEUWS, S.; MOLENBERGHS, G. The analysis of multivariate longitudinal data: A review. **Stat methods Med Res**, v. 23, n. 1, p. 42–59, 2014.

WESTFALL, J.A.; SCOTT, C.T. Taper models for commercial tree species in the northeastern United States. **Forest Science**, v. 56, n. 6, p. 515–528, 2010.

WESTFALL, J.A.; MCROBERTS, R.E.; RADTKE, P.J.; WEISKITTEL, A.R. Effects of uncertainty in upper-stem diameter information on tree volume estimates. **European Journal of Forest Research**, v. 135, n. 5, p. 937–947, 2016.

YANG, Y.; TITUS, S.J.; HUANG, S. Modeling individual tree mortality for white spruce in Alberta. **Ecological Modelling**, v. 163, n. 3, p. 209-222, 2003.

YANG, Y.; HUANG, S.; MENG, S.X. Development of a tree-specific stem profile model for white spruce: A nonlinear mixed model approach with a generalized covariance structure. **Forestry: An international Journal of Forest Research**, v. 82, n. 5, p. 541–555, 2009.

YAO, X.; TITUS, S.J.; MACDONALD, S.E. A generalized logistic model of individual tree mortality of aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. **Canadian Journal of Forest Research**, v. 31, n. 2, p. 283-291, 2001.

ZHANG, X.; CAO, Q.V.; DUAN, A.; ZHANG J. Modeling tree mortality in relation to climate, initial planting density, and competition in Chinese fir plantations using a Bayesian logistic multilevel method. **Canadian Journal of Forest Research**, v. 47, p. 1278-1285, 2017.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal Royal of Statistical Society. Series B**, v. 67, n. 2, p. 301-320, 2005.

APPENDIX

Scripts of each chapter and a digital version of the thesis is available on the following page https://github.com/luanfiorentin/thesis